The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus

Anonymous authors
Paper under double-blind review

Abstract

Explainable AI (XAI) is a rapidly evolving field that aims to improve transparency and trustworthiness of AI systems to humans. One of the unsolved challenges in XAI is estimating the performance of these explanation methods for neural networks, which has resulted in numerous competing metrics with little to no indication of which one is to be preferred. In this paper, to identify the most reliable evaluation method in a given explainability context, we propose MetaQuantus—a simple yet powerful framework that meta-evaluates two complementary performance characteristics of an evaluation method: its resilience to noise and reactivity to randomness. We demonstrate the effectiveness of our framework through a series of experiments, targeting various open questions in XAI, such as the selection of explanation methods and optimisation of hyperparameters of a given metric. We release our work under an open-source license: anonymous to serve as a development tool for XAI researchers and Machine Learning (ML) practitioners to verify and benchmark newly constructed metrics (i.e., "estimators" of explanation quality). With this work, we provide clear and theoretically-grounded guidance for building reliable evaluation methods, thus facilitating standardisation and reproducibility in the field of XAI.

1 Introduction

Since Explainable AI (XAI) is intended to increase trust and transparency in AI systems, it is necessary to evaluate the performance of proposed explanation methods to ensure their reliability. Apart from simpler or well-understood data domains where critical input features are known and models are interpretable (e.g., linear functions and shallow decision trees), in the context of more complex Machine Learning (ML) models such as neural networks (NNs), there is generally an absence of ground truth labels for explanations (Hedström et al., 2022). This makes it difficult to evaluate the performance of explanation methods since the exact outcomes of explanations oftentimes remain unknown and thus unverifiable (Dasgupta et al., 2022). Without consensus around how to define the quality or "correctness" of an explanation method, a variety of evaluation methods have been proposed. These efforts most commonly involve (i) measuring the extent to which desirable properties are fulfilled, e.g., through faithfulness or robustness analysis (Samek et al., 2017; Sundararajan et al., 2017; Montavon et al., 2018), (ii) generating well-defined, synthetic settings where explanation labels are simulated (Yang & Kim, 2019; Arras et al., 2022) or, (iii) evaluating explanations based on visual alignment with a human prior (Smilkov et al., 2017). Most relevant to our work is the first category of evaluation techniques or "metrics" whose goal is to estimate the quality of attribution-based explanations. We henceforth refer to these XAI evaluation methods as "quality estimators", or simply "estimators".

The abundance of explanation methods and an ever-growing number of quality estimators, combined with little guidance on how to use them, have caused confusion within the XAI and ML communities. Strong assertions of which explanation methods work and not (Adebayo et al., 2018; Sixt et al., 2020), followed by rebuttals (Sundararajan & Taly, 2018; Yona & Greenfeld, 2021; Binder et al., 2022), are ever-present. To answer the question of "which explanation method to use for a given task", we must first be able to define and measure the relevant qualities that an explanation method should fulfil. While preliminary efforts exist to address this issue (Brunke et al., 2020; Tomsett et al., 2020; Gevaert et al., 2022; Rong et al., 2022), to the best of our knowledge, there is currently no comprehensive solution that thoroughly evaluates the

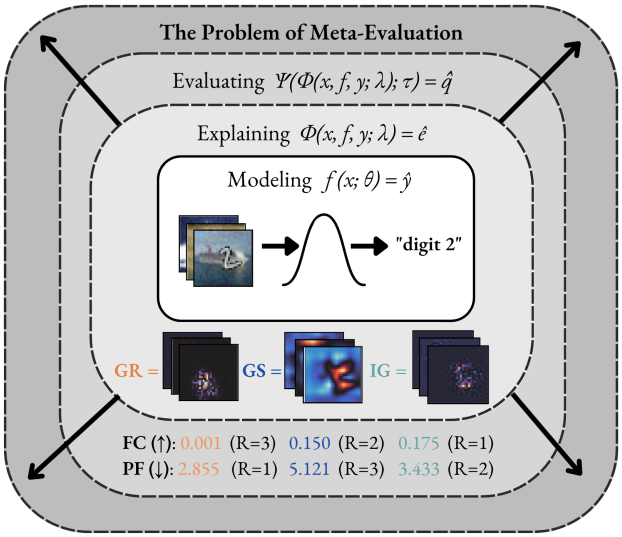


Figure 1: An illustration of the Problem of Meta-Evaluation through three phases: (i) Modeling, (ii) Explaining and (iii) Evaluating. (i) A ResNet9 model (He et al., 2016) is trained to classify digits from 0 to 9 on Customised-MNIST dataset (Bykov et al., 2021) (i.e., MNIST digits pasted on randomly sampled CIFAR-10 backgrounds). (ii) To understand the model's prediction, we use several explanation methods including *Gradient* (Mørch et al., 1995; Baehrens et al., 2010), *Integrated Gradients* (Sundararajan et al., 2017) and *GradientShap* (Lundberg & Lee, 2017), which are distinguished by their respective colours. (iii) To evaluate the quality of the explanations, we apply different estimators of faithfulness such as *Faithfulness Correlation* (FC) (Bhatt et al., 2020) and *Pixel-Flipping* (PF) (Bach et al., 2015), which return a correlation coefficient and an AUC score, respectively. However, since the scores vary depending on the estimator, both in range and direction, with lower or higher scores indicating more faithful explanations, interpreting the resulting faithfulness scores remains difficult for the practitioner.

various estimators used to compare, select and reject different explanation methods in XAI. Previous efforts at addressing this issue have been limited in scope and do not provide a thorough theoretical motivation. With this work, we aim to fill this critical yet largely neglected research gap.

In this work, we propose a solution to the problem of "meta-evaluation" in XAI, which we define as the process of evaluating the evaluation method (i.e., "quality estimator") itself. This problem arises as we select and quantitatively compare explanation methods for a given model, dataset and task. As illustrated in Figure 1, we can apply various estimators to compare the explanation methods' faithfulness, which measures how closely the explanations align with the predictive behaviour of the model (the experimental details are described in Appendix A.4). However, the estimators rank the same explanation methods differently, e.g., the *Gradient* method (Mørch et al., 1995; Baehrens et al., 2010) is both ranked the highest (R=1) and the lowest (R=3) depending on the estimator used. With a disagreement about which explanation method is superior (Krishna et al., 2022) coupled with little to no guidance on how to identify a high-quality estimator (Wang & Wang, 2022), practitioners may unknowingly choose an inferior quality estimator which ultimately results in a selection of an explanation method that presents a less faithful explanation to the end user.

To tackle the issue of explanation method selection, we propose a simple yet comprehensive framework called `MetaQuantus`, which primary purpose is to provide an objective, independent view of the estimator's performance by meta-evaluating it against two failure modes: resilience to noise (NR) and reactivity to adversary (AR). Similar to how software systems undergo vulnerability- and penetration tests before getting deployed in a larger system, we apply this framework to stress test the estimators. If vulnerabilities in the quality estimator are discovered, e.g., high sensitivity to noise in input or low reactivity to randomness, appropriate actions can be taken to improve the estimators. The contribution of this work is three-fold.

- First, we provide a clear argument for why performance evaluation of XAI methods is challenging (Section 2.2) including which variables in the evaluation process are verifiable vs. unverifiable.

- Second, based on these findings, we propose a framework to meta-evaluate quality estimators' performance (Section 3), with sound theoretical underpinning and wide applicability across various data-, models-, explanation methods- and metric domains.
- Third, we experimentally demonstrate that the meta-evaluation framework can solve a variety of XAI-related tasks, e.g., selecting a metric in a given category of explanation quality and optimising a metric's hyperparameters. We conduct a series of experiments on a variety of SOTA explanation methods, datasets and models and consequently, generate novel insights into the behaviour of the estimators (Section 6).

We find it important to point out that we have no interest in developing yet another evaluation procedure or presenting an additional view of explanation quality. The real need in our community lies in developing standardised tools to validate the quality estimators that already exist. It is surprising to us that very little effort has so far been directed towards this important area of analysing the behaviour of estimators. With this work, we hope to provide more clarity and guidance on how to effectively evaluate explanation methods and moreover, help with the selection process of choosing a quality estimator in a given explainability context.

1.1 Related Works

Despite much activity towards the development of estimators to assess explanation quality, e.g., (Bach et al., 2015; Sundararajan et al., 2017; Bhatt et al., 2020; Nguyen & Martínez, 2020; Rieger & Hansen, 2020; Arias-Duart et al., 2021), limited attention has thus far been given to evaluating the estimators themselves. Only recently, increased attention has been raised on the intricacies that come with XAI evaluation, for example, the contributions of Brunke et al. (2020); Brocki & Chung (2022); Rong et al. (2022) emphasise the difficulty that comes with parameterising estimators. Another issue with evaluation was brought to light by Neely et al. (2021); Krishna et al. (2022), which revealed that explanations frequently disagree in their ranking of features. Additionally, several independent research groups were able to identify empirical "confounders" (Sundararajan & Taly, 2018; Kokhlikyan et al., 2021; Yona & Greenfeld, 2021; Binder et al., 2022) affecting the well-adopted *Model Parameter Randomisation* test (Adebayo et al., 2018). From these publications, it seems worryingly "easy to get it wrong" when it comes to evaluating explainable methods empirically. There still remains a lot of ambiguity when it comes to determining what makes up a good or bad evaluation metric (Wang & Wang, 2022).

Within the scope of evaluating quality estimators, preliminary efforts exist but a unified effort is required. Since there is little to no consensus on how to determine the true value of a metric—when new metrics are introduced, they are oftentimes assessed based on single perspectives such as by randomisation experiments (Rieger & Hansen, 2020; Arias-Duart et al., 2021) or ranking consistency (Rong et al., 2022). All of these mentioned works are undoubtedly steps in the right direction, but what is missing is a broader, more comprehensive framing of what an evaluation method ought to fulfil. With this work, we aim to fill this gap.

2 Preliminaries

In the following, we derive a mathematical definition of the evaluation problem in XAI by outlining the key elements required to perform quality estimation on a given explanation method. In the succeeding section, we discuss the Challenge of Unverifiability (CoU) which explains why meta-evaluation is theoretically difficult. All notation used throughout this paper can be found in the Appendix A.7.

2.1 The Evaluation Problem

Consider a supervised classification problem[1] where we have a black-box model f parameterised by θ that has been trained on a given training dataset $\boldsymbol{X}_{\text{tr}} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ to map an input $\boldsymbol{x} \in \mathbb{R}^D$ to an output class $y \in \{1, \ldots, C\}$, with a trained functional mapping such as:

$$f(\boldsymbol{x}; \theta) = \hat{y}, \tag{1}$$

[1]Since classification tasks are commonly encountered in the XAI community, it is chosen to illustrate the Evaluation Problem. However, as discussed in Appendix A.1.1, our statements also apply to other prediction scenarios.

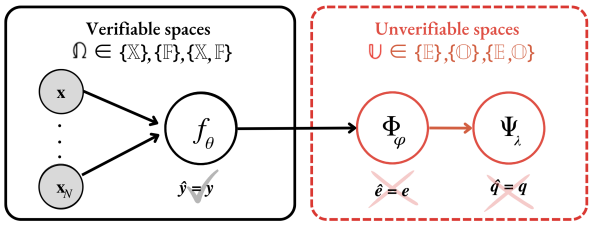


Figure 2: A visual representation of the conditional dependencies between variables in XAI evaluation. The information flows from modelling to explaining and evaluating the explanations, i.e., $\Psi \circ \Phi \circ f$, which is indicated by the direction of the arrows in the directed acyclic graph (DAG). The colours indicate if the spaces have verifiable (black) or unverifiable outcomes (red).

More generally, we can define the model function $f : \mathbb{X} \mapsto \mathbb{Y}$ that maps inputs from the instance space $\mathbb{X}$ to predictions in the label space $\mathbb{Y}$ with $\boldsymbol{x} \in \mathbb{X}$ and $\hat{y} \in \mathbb{Y}$. Let $\mathbb{F}$ denote the function space such that $f \in \mathbb{F}$. To quantitatively estimate the performance of model f, we compute the prediction error on a given test dataset $\boldsymbol{X}_{\text{te}}$ where there exists a label y for each prediction $\hat{y}$. To understand the reasoning of the model f behind a certain prediction $\hat{y}$, we can apply one of the many proposed *local* explanation methods (Smilkov et al., 2017; Zeiler & Fergus, 2014; Sundararajan et al., 2017; Selvaraju et al., 2020; Bykov et al., 2021) as follows:

$$\Phi(\boldsymbol{x}, f, \hat{y}; \lambda) = \hat{\boldsymbol{e}}, \tag{2}$$

where $\Phi : \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}^D$ is an explanation function that is parameterised by λ and which distributes attributions to each individual feature in $\boldsymbol{x}$ according to its importance, typically visualised in an explanation map $\hat{\boldsymbol{e}} \in \mathbb{R}^D$. Let $\mathbb{E}$ denote the space of possible explanations such that $\Phi \in \mathbb{E}$.

Similar to how we compute the prediction error to estimate the performance of a model f, to evaluate the quality of the explanation function Φ, we compute the explanation error, requiring a ground truth explanation $\boldsymbol{e}$. These labels are, however, generally not available for complex ML models and in particular NNs, since their inner workings are considered unknown (Bellido & Fiesler, 1993; Benítez et al., 1997; Dasgupta et al., 2022). Therefore, XAI researchers and ML practitioners must resort to indirect approaches to estimate the quality of a given explanation, e.g., by measuring the explanation's relative fulfilment of certain human-defined properties. Recent work by Hedström et al. (2022) has proposed to group these properties of explanation quality into six categories; (a) faithfulness, (b) robustness, (c) localisation, (d) randomisation, (e) complexity and (f) axiomatic metrics which provide a natural framework to compare and analyse explanation quality. A summary of these explanation categories can be found in Appendix A.2 (see Equations 11-15).

We provide a generalised notation for quality estimation of attribution-based explanation methods as follows. Let $\Psi_\tau : \mathbb{E} \times \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}$ be a quality estimator that is parameterised by τ and takes one explanation and returns one scalar value ("quality estimate") to indicate the quality of the explanation. The evaluation of an explanation, i.e., quality estimation, can be written as follows:

$$\Psi(\Phi, \boldsymbol{x}, f, \hat{y}; \tau) = \hat{q}, \tag{3}$$

where Ψ represents the quality estimator and the whole space of possible estimators is denoted $\Psi \in \mathbb{O}$. Mathematical descriptions of such estimators can be found in Appendix (see Equations A.3).

2.2 The Challenge of Unverifiability

The goal of quantitative evaluation is to provide an objective measure of the quality of an explanation. However, due to missing ground truth, the quantitative assessment of neural network explanations remains non-trivial. To clarify where this difficulty arises, we represent the process of XAI evaluation as a directed acyclic graph (DAG), as seen in Figure 2. Here, each node represents a random variable and the edges represent the relationships between the variables, with uncertainty of a parent node propagating to its child node. We separate the nodes between verifiable- and unverifiable spaces. The verifiable spaces are spaces where ground truth labels are available, i.e., $\mathbb{O} \in \{\{\mathbb{X}\}, \{\mathbb{F}\}, \{\mathbb{X}, \mathbb{F}\}\}$ and the unverifiable spaces include spaces where there is an absence of labels, i.e., $\mathbb{U} \in \{\{\mathbb{E}\}, \{\mathbb{O}\}, \{\mathbb{E}, \mathbb{O}\}\}$.

As indicated by the direction of the arrows, a key observation is that in XAI evaluation there exists a conditional dependency between the variables of modelling, explaining and evaluating (the explanations). This further means that since the evaluation function is applied to the results of the unverifiable explanation function, the evaluation outcome also renders unverifiable. We refer to this phenomenon as the Challenge of Unverifiability. Another key observation is that we cannot determine the accuracy or validity of an estimator (i.e., whether it actually measures the intended quality) since such assessment requires access to ground truth labels. However, as reliability analysis does not depend on the availability of ground truth labels, it is still possible to study the reliability of an estimator, which refers to its overall consistency ("does this estimator produce similar results under consistent conditions?"). This can be achieved by repeatedly measuring the evaluation outcomes that result from fixing the unverifiable parameters and functions and only varying the elements of the verifiable spaces. In the following, we will use the distinction between verifiable- and unverifiable spaces to systematically and controllably measure the performance of quality estimators.

3 A Meta-Evaluation Framework

While the Challenge of Unverifiability makes meta-evaluation of quality estimators challenging, it is still possible to study the performance characteristics of an estimator through the lens of reliability. To this end, we developed a three-step framework, which is a higher-level evaluation scheme that examines quality estimators that have themselves been used to evaluate a particular explanation method.

3.1 Defining Failure Modes

Without ground truth information, we cannot validate or optimise the quality estimators against what we want them to fulfil, but we can instead articulate edge-case scenarios or behaviours that we do not want them to exhibit. For this purpose, we formulate failure modes which are described in the following.

Failure Mode 1 (Noise Resilience)**.** *A quality estimator should be resilient to minor perturbations of its input parameters.*

Similar to the robustness property of explanation functions and specifically *Lipschitz Continuity* (Alvarez-Melis & Jaakkola, 2018; Montavon et al., 2018; Yeh et al., 2019), where small changes in the input should only lead to small changes in the explanation, noise resilience (NR) evaluates the extent to which a quality estimator is robust towards minor perturbations of its inputs. Following our general perturbation Definition 3 in Appendix A.2, we define a minor perturbation $\mathcal{P}_{\mathbb{\Omega}}^{M}$ of any verifiable space $\mathbb{\Omega}$ as follows:

Definition 1 (Minor Perturbation)**.** Let $\mathcal{P}_{\mathbb{\Omega}}(\boldsymbol{\omega})$ be a perturbation function of $\boldsymbol{\omega} \in \mathbb{\Omega}$, $\hat{y} = f(\boldsymbol{x};\theta)$ be the original prediction of the network and y' be the prediction after the perturbation. Then $\mathcal{P}_{\mathbb{\Omega}}(\boldsymbol{\omega})$ is minor $\mathcal{P}_{\mathbb{\Omega}}^{M}$, if $\forall\ \hat{y},\ y' \in \{\{f(\mathcal{P}_{\mathbb{X}}^{M}(\boldsymbol{x});\theta)\}, \{f(\boldsymbol{x};\mathcal{P}_{\mathbb{F}}^{M}(\theta))\}, \{f(\mathcal{P}_{\mathbb{X}}^{M}(\boldsymbol{x});\mathcal{P}_{\mathbb{F}}^{M}(\theta))\}\},\ \exists\ \epsilon \in \mathbb{R}\ \epsilon \ll 1$ such that:

$$||\hat{y} - y'||_p \leq \epsilon$$

For classification, we employ L1-norm with $p = 1$, thus, Definition 1 states that the predicted label y' stays unchanged after the perturbation, i.e., $\hat{y} \approx y'$. Similar to works by Brunke et al. (2020); Brocki & Chung (2022); Rong et al. (2022), we measure the vulnerability of quality estimators to variations or "minor confounds" in the estimator. However, in contrast to these aforementioned works, we only perturb in the verifiable space by means of measuring the change in the model decision on a sample before and after perturbation, and thus we can control and directly measure the strength of the perturbation. Accordingly, to quantitatively examine Failure Mode 1, we expose the estimator to perturbations with small or minor impacts. Complementary to testing an estimator's resilience to noise, we also formulate a second failure mode to test whether a quality estimator produces a significant change when exposed to disruptive perturbation, i.e., randomisation to any of its inputs.

Failure Mode 2 (Adversary Reactivity)**.** *A quality estimator should be reactive to disruptive perturbations of its input parameters.*

Previous research has noted that the estimators' scores should be conceivably different when produced for a random explanation (Rieger & Hansen, 2020) or a randomly initialised model (Arias-Duart et al., 2021). Our

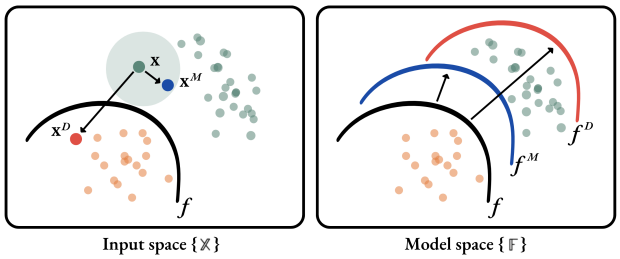


Figure 3: An illustration of minor versus disruptive perturbation in the different spaces (left: $\mathbb{X}$, right: $\mathbb{F}$) for a classification task. The direction of the arrows shows how the respective perturbations are realised, where blue and red colours indicate a minor- or disruptive perturbation, respectively. The minor perturbation keeps the decision boundary intact, either by perturbing a sample $\boldsymbol{x}^M$ (left) or perturbing the model itself f^D (right). The disruptive perturbation implies that the decision boundary is crossed either through a sample $\boldsymbol{x}^D$ (left) or model f^D (right).

approach is similar in that it also seeks to disrupt the explanation process. However, since we can control the perturbation strength in the verifiable spaces, we can make more well-grounded claims about the expected outcomes of a perturbation. Theoretically, we define disruptive perturbations $\mathcal{P}^D_{\mathbb{\Omega}}$ contrary to Definition 1.

Definition 2 (Disruptive Perturbation)**.** $\mathcal{P}_{\mathbb{\Omega}}(\boldsymbol{\omega})$ be a perturbation function of $\boldsymbol{\omega} \in \mathbb{\Omega}$, $\hat{y} = f(\boldsymbol{x};\theta)$ be the original prediction of the network and y' be the prediction after the perturbation. Then $\mathcal{P}_{\mathbb{\Omega}}(\omega)$ is disruptive $\mathcal{P}^D_{\mathbb{\Omega}}$, if $\forall\ \hat{y},\ y' \in \{\{f(\mathcal{P}^D_{\mathbb{X}}(\boldsymbol{x});\theta)\}, \{f(\boldsymbol{x};\mathcal{P}^D_{\mathbb{F}}(\theta))\}, \{f(\mathcal{P}^D_{\mathbb{X}}(\boldsymbol{x});\mathcal{P}^D_{\mathbb{F}}(\theta))\}\}\ \exists\ \epsilon \in \mathbb{R},\ \epsilon \ll 1$ such that:

$$||\hat{y} - y'||_p > \epsilon.$$

In a classification context, Definition 2 implies a change in the predicted class label. Figure 3 illustrates the main difference between minor and disruptive perturbations, which is that the decision boundary remains uncrossed or crossed, respectively. In Appendix A.1.1, we expand the Definitions 1 and 2 to other problem settings such as multi-label classification and also discuss how adversarial attacks relate to these definitions.

Using Definitions 1 or 2, we can generate perturbed quality estimates q' by applying a minor or disruptive perturbation on the verifiable spaces in the input, model, or input- and model spaces simultaneously:

$$\hat{q} \in \{\ \Psi(\Phi, \mathcal{P}^t_{\mathbb{X}}(\boldsymbol{x}), f, \hat{y}),\ \ \Psi(\Phi, \boldsymbol{x}, \mathcal{P}^t_{\mathbb{F}}(\theta), \hat{y})),\ \ \Psi(\Phi, \mathcal{P}^t_{\mathbb{X}}(\boldsymbol{x}), \mathcal{P}^t_{\mathbb{F}}(\theta), \hat{y}))\ \},\tag{4}$$

where the superscript of the perturbation function, $t \in \{M, D\}$ indicates the perturbation strength. For simplicity, we omit the hyperparameters τ, λ from Equation 4. By repeating this perturbation (Equation 4) multiple times, we gather sets of perturbed estimates for meta-evaluation analysis. In the next section, we provide a detailed description of how this analysis is performed.

3.2 Formulating Consistency Criteria

To determine whether a quality estimator appropriately circumvented a failure mode, we can measure the similarity of its quality estimates before and after the perturbation. After a minor perturbation, when testing for noise resilience, we would expect that the scores are similarly distributed. Conversely, for disruptive perturbations, when testing for reactivity to adversary, we would anticipate a large response to information annihilation of the explanation process by means of scores being dis-similarly distributed. We formalise this idea in our **I**ntr**a**-**C**onsistency (IAC) criterion as follows:

$$\mathbf{IAC} = \frac{1}{K}\sum_{k=1}^{K} d(\hat{\boldsymbol{q}}, \boldsymbol{q}'_k),\tag{5}$$

where $\hat{\boldsymbol{q}}$ refers to unperturbed estimates and $\boldsymbol{q}'_k \in \mathbb{R}^N, k = (1, \ldots, K)$ is a set of perturbed quality estimates, replicated K times for N test samples (see Equation 4) such that $\boldsymbol{Q} = [\boldsymbol{q}'_1, \ldots, \boldsymbol{q}'_K] \in \mathbb{R}^{N\times K}$. Here, d refers to a statistical significance measure $d : \mathbb{R}^N \times \mathbb{R}^N \mapsto \mathbb{R}$ that takes a set of unperturbed- and perturbed estimates

and returns a p-value. A high p-value indicates that $\hat{\boldsymbol{q}}$ and $\boldsymbol{q}'_k$ are similarly distributed and a low p-value value means that the estimates are differently distributed. Accordingly, Equation 5 returns the average p-value across all perturbed samples over K perturbations, with IAC $\in [0, 1]$. Since the nominal values of quality estimators can vary and often have little to no semantic meaning, we use the non-parametric *Wilcoxon signed-rank test* (Wilcoxon, 1945) which does not carry strong assumptions about the data distribution, only about its ranking. In addition to the intra-consistency analysis, we also measure whether quality estimators exhibit consistent behaviour in terms of ranking. This type of inter-consistency analysis is commonly used in Explainable AI research (Tomsett et al., 2020; Gevaert et al., 2022; Hedström et al., 2022; Rong et al., 2022) and complements the aforementioned by involving more than one explanation method. Let $\bar{\boldsymbol{Q}} \in \mathbb{R}^{N \times L}$ denote a matrix for the unperturbed estimates $\hat{\boldsymbol{q}}$ for L explanation methods and $\bar{\boldsymbol{Q}}' \in \mathbb{R}^{N \times L}$ be a matrix for the perturbed estimates $\boldsymbol{q}'_k$, which are both averaged over K perturbations. We formulate the **I**nt**e**r-**C**onsistency (IEC) criterion as follows:

$$\textbf{IEC} = \frac{1}{N \times L} \sum_{i=1}^{N} \sum_{j=1}^{L} U_{i,j}^{t} \tag{6}$$

where $U_{i,j}^{t} \in [0, 1]$ are entries of a binary ranking agreement matrix $\boldsymbol{U}$ that takes quality estimates from $\bar{\boldsymbol{Q}}$ and $\bar{\boldsymbol{Q}}'$ and populates the entries according to the interpretation of ranking. Here, IEC $= 1$ indicates perfect ranking consistency and IEC $= 0$ the absence of it, where IEC $\in [0, 1]$. The perturbation strength is indicated in the superscript $t \in \{M, D\}$. The interpretation of ranking is different depending on the perturbation strength, i.e., minor or disruptive. For minor perturbations, we measure if the quality estimator ranks different explanation methods similarly. We define U^M for minor perturbations with entries such as:

$$U_{i,j}^{M} = \begin{cases} 1 & \bar{r}_j^M = \bar{r}_j \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\bar{\boldsymbol{r}}^M = r(\bar{\boldsymbol{Q}}_{i,:}^M)$ with $\bar{\boldsymbol{Q}}^M := \bar{\boldsymbol{Q}}'$ and $\bar{\boldsymbol{r}} = r(\bar{\boldsymbol{Q}}_{i,:})$ are ranking vectors given a ranking measure $r : \mathbb{R}^L \mapsto \mathbb{R}^L$ that takes each row in $\bar{\boldsymbol{Q}}_{i,:}^M$ and $\bar{\boldsymbol{Q}}_{i,:}$, respectively and sorts the values in descending order. Each entry $\bar{r}_j^M \in \mathbb{N}$ corresponds to integers indicating their relative rank. For example, suppose we have one sample $\boldsymbol{x}$, three explanation methods and their corresponding quality estimates, such as $\bar{\boldsymbol{Q}}_{i,:}^M = [0.76, 0.86, 0.66]$. Then the results obtained from applying r would be $\bar{\boldsymbol{r}}^M = [2, 1, 3]$. An optimally-performing estimator would provide the same rankings for $\bar{\boldsymbol{r}}^M$ as $\bar{\boldsymbol{r}}$ for all N inputs, resulting in IEC $= 1$. However, as discussed in Section 6, the reality is that many estimators often conflict with the optimal.

For disruptive perturbations, we interpret ranking consistency differently. Here, as explained in-depth in Appendix A.1.2, we measure how consistently the quality estimator ranks estimates from $\bar{\boldsymbol{Q}}$ higher than $\bar{\boldsymbol{Q}}^D := \bar{\boldsymbol{Q}}'$. We define $\boldsymbol{U}^D$ for disruptive perturbations with entries such as:

$$U_{i,j}^{D} = \begin{cases} 1 & \bar{Q}_{i,j}^D < \bar{Q}_{i,j} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where the quality estimates $\bar{Q}_{i,j}^D$ are generated for an explanation with respect to the same class as the one predicted for the unperturbed estimate $\bar{Q}_{i,j}$. For some estimators, e.g., in the robustness category, lower values are considered better than higher values, for which we invert the comparison symbol in Equation 8.

3.3 Quantifying Meta-Consistency

To conclude the framework, we want to characterise the performance of a quality estimator with a single Meta-Consistency (MC) score. To capture both the estimator's resilience to noise (NR) and its reactiveness to adversary (AR), we average over the two criteria for both failure modes:

$$\textbf{MC} = \left(\frac{1}{|\boldsymbol{m}^*|} \right) \boldsymbol{m}^{*T} \boldsymbol{m} \quad \text{where} \quad \boldsymbol{m} = \begin{bmatrix} \textbf{IAC}_{NR} \\ \textbf{IAC}_{AR} \\ \textbf{IEC}_{NR} \\ \textbf{IEC}_{AR} \end{bmatrix} \tag{9}$$

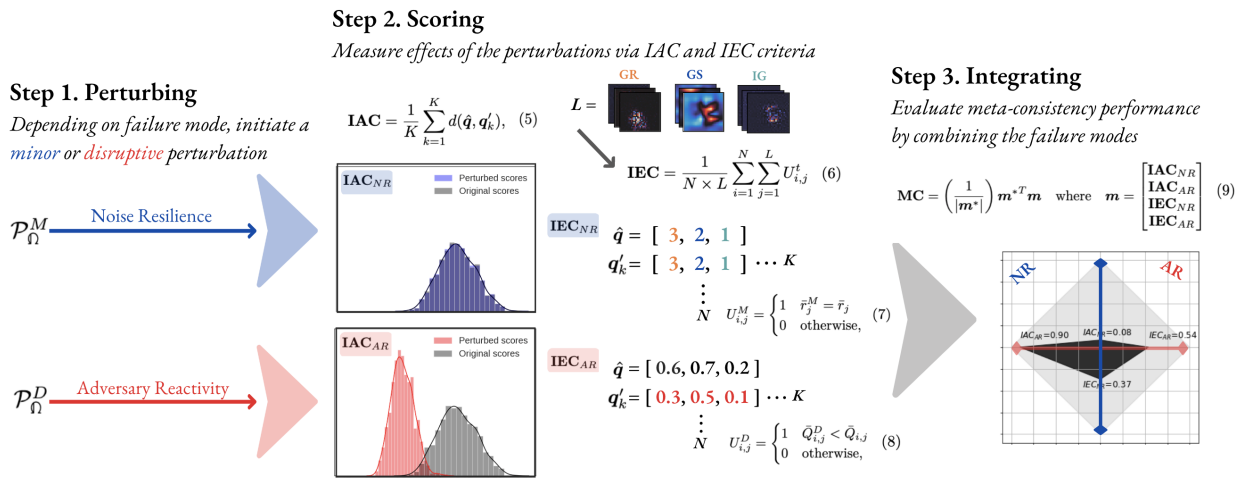


Figure 4: Meta-evaluation of quality estimators is performed in three steps: (i) Perturbing, (ii) Scoring and (iii) Integrating. (i) First, a minor or disruptive perturbation is induced depending on the failure mode, i.e., $\mathcal{P}_{\Omega}^{M}$ for NR and $\mathcal{P}_{\Omega}^{D}$ for AR. (ii) Second, the estimator's intra- and inter-consistency are calculated to assess each performance dimension. The IAC score captures the extent that the estimator produces similar or dis-similar scores with respect to $\hat{\boldsymbol{q}}$ and $\boldsymbol{q}'_k$, which is illustrated through the distribution plots, where for NR and AR, the score distributions are overlapping and non-overlapping, respectively. The IEC score expresses ranking consistency. NR measures how consistently the estimator ranks different explanation methods and AR calculates how consistently the perturbed scores are lower than the unperturbed scores. (iii) In the final step, we integrate the previous steps and produce an MC score that summarises the estimator's performance: its resilience to noise and reactivity to adversary.

and $\boldsymbol{m}^* = \mathbb{1}^4$ represents an optimally performing quality estimator as defined by the all-one indicator vector. A good quality estimator should produce an MC score close to 1 as higher values indicate better performance on the tested criteria[2], where MC $\in [0, 1]$. An estimator that demonstrates a balance of resilience against minor perturbations and reactiveness towards disruptive perturbations—as evidenced through its score distribution and ranking of different explanation methods—would achieve high meta-consistency scores with our framework. Our proposed score has the advantage of being both concise and comprehensive, as it provides a summary of the performance characteristics of an estimator while also taking into account multiple criteria. For a full overview of the framework, please see Figure 4.

4 Practical Evaluation

Within the framework of meta-evaluation, it is necessary to generate perturbed quality estimates for analysis. To accomplish this, we developed a series of practical tests. The methodology behind these tests is simple and thus easily extensible through the tests made available in the repository (`anonymous`). First, the space in which perturbations will be applied is selected, with options being either the input or the model. Second, based on the chosen space, an appropriate type of noise is defined. To ensure that the perturbations are meaningful and relevant to the task at hand, the noise type should be chosen contextually with respect to the data domain. For example, when perturbing the input space for images, we define a test as follows:

Input Perturbation Test (IPT). Apply i.i.d additive uniform noise such that $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \boldsymbol{\delta}_i$ with $\boldsymbol{\delta}_i \sim \mathcal{U}(\alpha, \beta)$ where for noise resilience, $\hat{\boldsymbol{x}}_i$ fulfills Definition 1 and for adversary reactivity, $\hat{\boldsymbol{x}}_i$ fulfills Definition 2

where α, β have to be chosen according to the data domain and respective failure mode (e.g., set $\alpha = -0.001, \beta = 0.001$ for NR and $\alpha = 0.0, \beta = 1.0$ for AR). To maintain the statistics of the data distribution, we clip α, β to the maximum and the minimum value of the test set, respectively. Moreover, when perturbing the model space, to maintain the variance of the network, we follow an established methodology by Bykov et al. (2021) and scale the learned weights $\boldsymbol{\theta}$ of the model f as follows:

[2]When computing intra-consistency scores for AR, we apply reverse scoring, i.e., $1 - \mathbf{IAC}_{AR}$, so that all elements in the meta-evaluation vector (Equation 9) can be interpreted in the same way, i.e., that higher values are better.

Model Perturbation Test (MPT). Apply multiplicative Gaussian noise to all weights of the network, i.e., $\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta} \cdot \boldsymbol{\nu}_i$ with $\boldsymbol{\nu}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = 1$ and for noise resilience, $\hat{\boldsymbol{\theta}}_i$ fulfills Definition 1 and for adversary reactivity, $\hat{\boldsymbol{\theta}}_i$ fulfills Definition 2

where for $\boldsymbol{\Sigma}$ to be consistent with either Definition 1 or 2, it is set based on the specific context of the model and task being considered (e.g., $\boldsymbol{\Sigma} = 0.001$ for NR and $\boldsymbol{\Sigma} = 2.0$ for AR). Third, to collect sets of perturbed quality estimates for intra- and inter-consistency analysis, we repeat the process of perturbation (as outlined in IPT and MPT) and subsequent evaluation (using Equation 4) under K runs. Finally, we compute the MC score. For sanity-checking experiments of the tests, see Appendix A.5. Moreover, as a third testing scenario, it is theoretically possible to perturb both the input- and model spaces simultaneously, i.e., $\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}), \mathcal{P}_{\mathbb{F}}(\theta)$ as well as their respective latent spaces. This we leave for future work.

5 Experimental Setup

In this section, we give a brief account of the experimental setup, including the datasets, models, explanation methods and estimators used in this work. Further details can be found in Appendix A.4.

In our experiments, we benchmark five different categories of explanation quality and within each category, we have selected two estimators as follows: *Complexity* (CO) (Bhatt et al., 2020), *Sparseness* (SP) (Chalasani et al., 2020), *Faithfulness Correlation* (FC) (Bhatt et al., 2020), *Pixel-Flipping* (PF) (Bach et al., 2015), *Max-Sensitivity* (MS) (Yeh et al., 2019), *Local Lipschitz Estimate* (LLE) (Alvarez-Melis & Jaakkola, 2018), *Pointing-Game* (PG) (Zhang et al., 2018), *Relevance Mass Accuracy* (RMA) (Arras et al., 2022), *Model Parameter Randomisation Test* (MPR) (Adebayo et al., 2018) and *Random Logit* (RL) (Sixt et al., 2020). Each estimator evaluates explanations from a popular category of post-hoc attribution methods, including both gradient-based- and model-agnostic techniques: *Gradient* (Mørch et al., 1995; Baehrens et al., 2010), *Saliency* (Mørch et al., 1995), *GradCAM* (Selvaraju et al., 2020), *Integrated Gradients* (Sundararajan et al., 2017), *Input×Gradient* (Shrikumar et al., 2016), *Occlusion* (Zeiler & Fergus, 2014) and *GradientSHAP* (Lundberg & Lee, 2017) from which we generate explanations with respect to a sample's predicted class. For comparability, we normalise the explanations by dividing the attribution map by the square root of its average second-moment estimate (Binder et al., 2022). The mathematical definitions of the estimators are described in Appendix A.3. For metric implementations, we use the `Quantus` library (Hedström et al., 2022).

We use four image classification datasets for our experiments: MNIST (LeCun et al., 2010), fMNIST (Xiao et al., 2017), customised-MNIST (i.e., cMINST) (Bykov et al., 2021) and ILSVRC-15 (i.e., ImageNet) (Russakovsky et al., 2015) and use different black-box NNs, including architectures such as LeNets (LeCun et al., 1998) and ResNets (He et al., 2016) which contributes to the robustness of our experimental findings.

6 Results

Many open questions remain in the field of XAI. In this section, we show how meta-evaluation can help bring clarity to a subset of those problems such as (i) estimator selection, (ii) optimising hyperparameters of an estimator and (iii) evaluating the category convergence, i.e., the extent that estimators within the same category of explanation quality measure the same concept. We prioritise the topic of metric selection in the main manuscript and provide a detailed analysis and discussion of the experiments addressing questions (ii) and (iii) in Appendix A.6. Instructions for how to reproduce the experiments can be found in the repository (`anonymous`).

6.1 Benchmarking

As a first example, we will demonstrate how meta-evaluation can be used to select a certain quality estimator for a given category of explanation quality. To this end, we set up a benchmarking experiment, where we take two popular estimators from five different explanation quality categories and evaluate six explanation methods $L = \{$*Gradient, Saliency, GradCAM, Integrated Gradients, Occlusion, GradientShap*$\}$ using `MetaQuantus`. Since the choice of L has a minimal influence on the MC scores (see experiments in Appendix A.5.2), we omit results from other tested sets of explanation methods in the main manuscript.

Table 1: Benchmarking results for MNIST dataset, aggregated over 3 iterations with $K = 5$. IPT results are in grey rows and MPT results are in white rows. The symbol $\overline{\text{MC}}$ denotes the averages of the MC scores over IPT and MPT. The top-performing method in each category, which outperforms the bottom-performing method by at least 2 standard deviations, is underlined. Higher values are preferred for all scoring criteria.

Category	**Estimator**	$\overline{\textbf{MC}}$ (↑)	**MC** (↑)	$\textbf{IAC}_{NR}$ (↑)	$\textbf{IAC}_{AR}$ (↑)	$\textbf{IEC}_{NR}$ (↑)	$\textbf{IEC}_{AR}$ (↑)
Complexity	Sparseness	$\underline{0.558 \pm 0.028}$	$\underline{0.640 \pm 0.043}$	0.209 ± 0.040	0.946 ± 0.086	0.837 ± 0.002	0.569 ± 0.046
			0.929 ± 0.063	0.053 ± 0.014	0.840 ± 0.005	0.084 ± 0.001	0.476 ± 0.013
	Complexity	0.521 ± 0.003	0.541 ± 0.007	0.009 ± 0.013	1.000 ± 0.000	1.000 ± 0.000	0.156 ± 0.014
			$\underline{0.500 \pm 0.000}$	0.167 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
Faithfulness	Faithfulness Corr.	0.540 ± 0.015	$\underline{0.537 \pm 0.003}$	0.477 ± 0.032	0.900 ± 0.023	0.190 ± 0.003	0.579 ± 0.008
			0.543 ± 0.026	0.500 ± 0.107	0.890 ± 0.005	0.190 ± 0.002	0.594 ± 0.005
	Pixel-Flipping	$\underline{0.626 \pm 0.039}$	$\underline{0.609 \pm 0.039}$	0.547 ± 0.139	0.963 ± 0.034	0.299 ± 0.001	0.626 ± 0.046
			0.644 ± 0.038	0.485 ± 0.141	1.000 ± 0.000	0.294 ± 0.006	0.796 ± 0.006
Localisation	Pointing-Game	$\underline{0.586 \pm 0.010}$	$\underline{0.672 \pm 0.020}$	0.977 ± 0.005	0.607 ± 0.075	0.996 ± 0.000	0.108 ± 0.012
			$\underline{0.500 \pm 0.000}$	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
	Relevance Rank Acc.	0.552 ± 0.015	0.613 ± 0.022	0.258 ± 0.062	0.793 ± 0.023	0.846 ± 0.001	0.553 ± 0.032
			0.491 ± 0.007	0.940 ± 0.019	0.071 ± 0.019	0.902 ± 0.003	0.051 ± 0.000
Randomisation	Random Logit	$\underline{0.666 \pm 0.004}$	$\underline{0.712 \pm 0.008}$	0.360 ± 0.041	0.969 ± 0.010	0.937 ± 0.003	0.581 ± 0.006
			$\underline{0.620 \pm 0.000}$	0.186 ± 0.000	0.874 ± 0.000	0.860 ± 0.000	0.562 ± 0.000
	Model Param. Rand.	0.583 ± 0.007	0.624 ± 0.005	0.264 ± 0.019	0.959 ± 0.000	0.764 ± 0.002	0.510 ± 0.001
			0.542 ± 0.010	0.250 ± 0.065	0.806 ± 0.028	0.647 ± 0.003	0.463 ± 0.004
Robustness	Max-Sensitivity	0.649 ± 0.007	0.754 ± 0.002	0.547 ± 0.064	0.938 ± 0.033	0.804 ± 0.001	0.726 ± 0.038
			0.545 ± 0.012	0.361 ± 0.053	1.000 ± 0.000	0.806 ± 0.005	0.011 ± 0.001
	Local Lipschitz Est.	$\underline{0.741 \pm 0.030}$	0.726 ± 0.026	0.484 ± 0.091	0.935 ± 0.088	0.736 ± 0.002	0.750 ± 0.034
			$\underline{0.756 \pm 0.034}$	0.519 ± 0.118	0.974 ± 0.017	0.740 ± 0.005	0.789 ± 0.006

6.2 Comparison of Estimators

The results are summarised in Table 1. The grey rows indicate the results from the Input Perturbation Test and the white rows show the results from the Model Perturbation Test. A more detailed discussion of the results, including additional datasets, can be found in Appendix A.6.3. From Table 1, we can observe that no tested estimator performs optimally, i.e., $\forall$ MC < 1. From column $\overline{\text{MC}}$, which displays the averaged MC scores (over IPT and MPT) we note that *Sparseness*, *Pixel-Flipping*, *Pointing-Game*, *Random Logit* and *Local Lipschitz Estimate* are the best-performing estimators in their respective category. From Figure 5 (right), we can observe that this comparison of MC scores is consistent across the tested datasets, which contributes to the generalisability of our findings. We further discuss the consistency of each estimator's rank (top or bottom) in Appendix A.6.4.

6.3 Comparison of Categories

The meta-evaluation framework can moreover be applied to gain insights into the performance characteristics of different estimators on a category-by-category basis. For this purpose, we represent the entries of the meta-evaluation vector as coordinates on a 2D plane and visualise the results as an area graph (see Figure 6). By inspecting the coloured areas of the respective estimators in terms of their size and shape, we can deduce the overall performance of both failure modes. Here, larger coloured areas imply better performance on the different scoring criteria and the grey area indicates the area of an optimally performing quality estimator, i.e., $\boldsymbol{m}^* = \mathbb{1}^4$. Each column of estimators represents a category of explanation quality, from left to right: *Complexity*, *Faithfulness*, *Localisation*, *Randomisation* and *Robustness*, which colour scheme we apply consistently across all figures.

As seen in Figure 6 (third column), the localisation estimators exhibit a notable deficiency in terms of adversary reactivity on the Input Perturbation Test. A low IPT score for adversary reactivity means that the estimators are insensitive to disruptive input perturbations, as evidenced by similar score distributions (low IAC) and an inability to rank disruptively perturbed explanations lower than unperturbed explanations (low IEC). Based on the definitions of these estimators (described in Equations 20-21), which include the *Pointing-Game* method (Zhang et al., 2018), which evaluates explanations by verifying that the highest attributed feature intersects with a given segmentation mask and the *Relevance Mass Accuracy* method (Arras et al., 2022), which calculates the amount of explainable mass intersecting with the segmentation mask—we would expect that these estimators perform well on this test since disrupted input usually leads

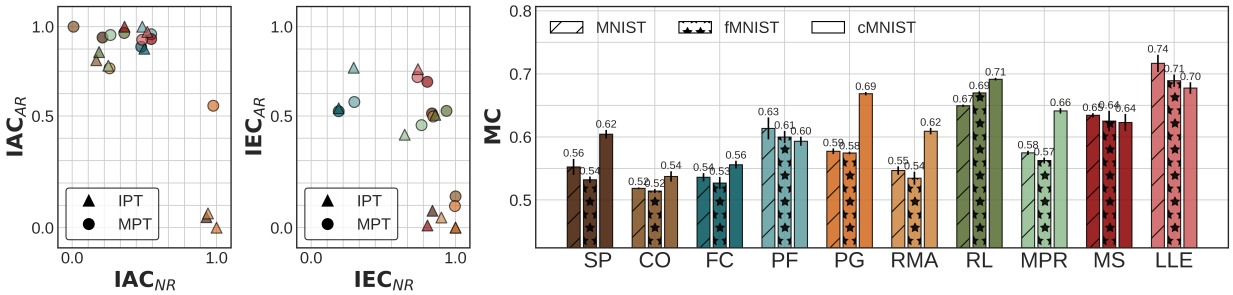


Figure 5: *Left:* A visualisation of the benchmarking results (Table 1), in particular IAC and IEC scores for noise resilience (x-axes) and adverse reactivity (y-axes). The colours indicate the estimator and the symbols show the test type, i.e., IPT and MPT, respectively. *Right:* A comparison of averaged meta-consistency performance for different quality estimators using MPT and IPT, aggregated over 3 iterations with $K = 5$, across different models {*LeNet, ResNet*} and datasets {*MNIST, fMNIST, cMNIST*}. Higher values are preferred.

scattered attributions. It may seem counterintuitive that these estimators lack reactivity to disruption, however, we posit that the reason for the poor reactivity to adversary is the estimators' inherent dependency on the segmentation mask. If the segmentation mask (relative to the object of interest, or the input) is large enough, high localisation scores are attainable irrespective of the "quality" of the explanation (Kohlbrenner et al., 2020). This finding is further validated by the increase in MC scores for cMNIST dataset, which has a smaller bounding box compared to MNIST and fMNIST (see details in Appendix A.4), where disruption evidently has an effect, as evidenced by higher AR scores, depicted in Figure 5 (right). Practitioners should be aware of this category's reliance (or oversensitivity) on the segmentation mask where relying solely or too heavily on this category in XAI evaluation may not be advisable.

The highest overall scores are obtained by the robustness and randomisation categories, which can be observed by their respective areas in Figure 6. One potential explanation for this is that the estimators in these categories already include some element of stochasticity in their metric definitions (see Equations 18-19 and 22-23, respectively) which may make them more resilient as well as reactive to perturbations. For example, both robustness estimators, i.e., *Max-Sensitivity* (Yeh et al., 2019) and *Local Lipschitz Estimate* (Alvarez-Melis & Jaakkola, 2018) rely on Monte Carlo sampling-based approximation where explanation methods are evaluated by examining their response to minor perturbation of the input, aggregated over multiple runs. In the randomisation category, the *Model Parameter Randomisation Test* (Adebayo et al., 2018) evaluates explanations by increasingly perturbing the model weights and *Random Logit* (Sixt et al., 2020) evaluates explanations by a random selection of an explanation of a non-target class. The complexity category has the lowest overall MC scores, which includes estimators such as *Sparseness* (Chalasani et al., 2020) and *Complexity* (Bhatt et al., 2020) that evaluate explanations by calculating their Gini coefficient and Shannon entropy, respectively. Given the simplicity of these calculations, this outcome is not surprising.

Another notable category of poor performance that is picked up by the meta-evaluation tests is faithfulness. Our results, which show a lack of resilience to noise in the ranking of explanation methods (low IEC), corroborate previous studies (Brunke et al., 2020; Brocki & Chung, 2022; Rong et al., 2022) that found that faithfulness metrics may rank explanation methods inconsistently when subjected to perturbation, such as changing the masking pixel strategy from, for example, "uniform" to "black". This trend is particularly evident in Figure 5 (left), where the points belonging to the faithfulness category have notably lower IEC scores compared to the other categories of explanation quality. A possible explanation is their well-documented sensitivity to parameterisation (Tomsett et al., 2020; Hedström et al., 2022; Rong et al., 2022).

While certain estimator categories, such as faithfulness, may present challenges such as parameterisation, it is not advisable to disregard their evaluation. Compared to categories such as complexity which are well-defined and simple to calculate, they may not offer as much information as categories such as faithfulness, which can provide important insights into how the explanation- and model functions are related. Relying on only one category to estimate explanation quality is therefore not recommended. This is especially true since an explanation function may trade one category of explanation quality over another (Bhatt et al., 2020),

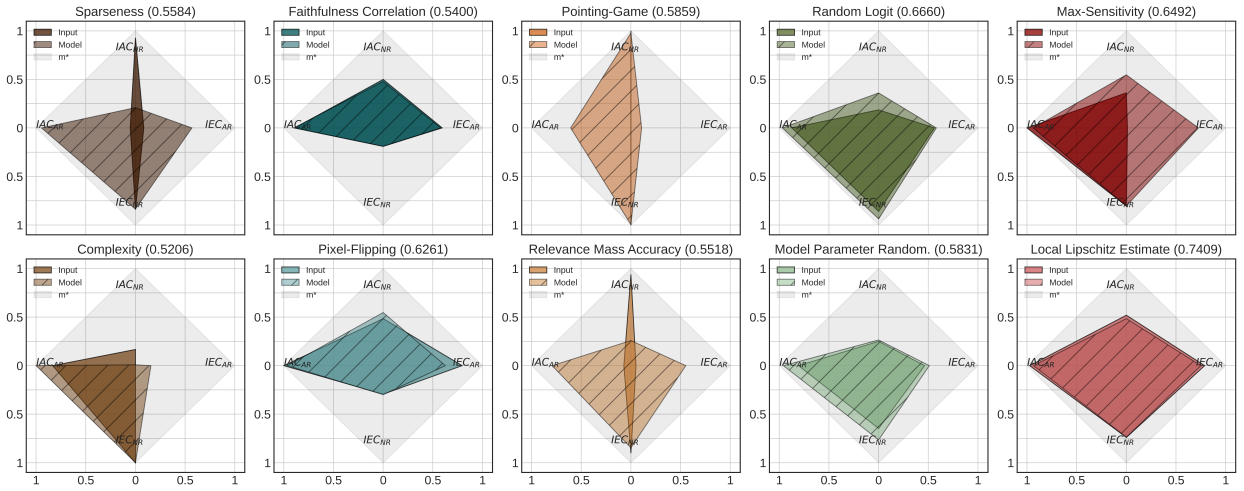


Figure 6: A graphical representation of the benchmarking results (Table 1), aggregated over 3 iterations with $K = 5$. Each column corresponds to a category of explanation quality, from left to right: *Complexity*, *Faithfulness*, *Localisation*, *Randomisation* and *Robustness*. The grey area indicates the area of an optimally performing estimator, i.e., $\mathbf{m}^* = \mathbb{1}^4$. The MC score (indicated in brackets) is averaged over MPT and IPT. Higher values are preferred.

for example, an explanation that is faithful may be too complex for the user to understand. Therefore, to avoid arriving at incomplete or incorrect conclusions about which explanation methods work (and not), it is of utmost importance for practitioners to approach evaluation by incorporating multiple definitions of explanation quality.

7 Conclusion

When we neither understand the general behaviour of the explanation methods nor the metrics that we apply to estimate their quality, we are bound to make mistakes. This problem in XAI is exacerbated by the fact that different estimators within the same category of explanation quality may rank the same explanation method differently (Neely et al., 2021; Krishna et al., 2022). Without an understanding of the performance characteristics of the estimators we employ, we risk presenting inferior explanation methods to the end user.

To address the problem of meta-evaluation, we propose a novel framework for identifying reliable metrics for XAI evaluation. We circumvent the Challenge of Unverifiability by evaluating the estimators through the lens of reliability—through perturbing the verifiable variables of XAI evaluation and thereafter analysing the estimator's outcomes under different failure modes, we can get an objective and independent characterisation of its performance. We show, in a series of experiments, how to use the framework for metric selection and for systematic evaluation of the strengths and weaknesses of individual metrics as well as general categories of explanation quality. Our findings show that (i) localisation estimators demonstrate a deficiency in terms of adversary reactivity, possibly due to their dependency on the segmentation mask and moreover that (ii) faithfulness category is inconsistent in its ranking and that (iii) randomisation and robustness are the highest-performing categories. It is advised that practitioners in the field of XAI take into account the limitations of various estimator categories and exercise caution when relying heavily on certain categories.

Evaluating the intrinsic value or "validity" of a quality estimator is, however, still an open and important question to consider. It is essential to keep in mind that the reliability of an estimator does not necessarily imply any intrinsic validity, e.g., an estimator's theoretical soundness (Binder et al., 2022). Moreover, since most explanation methods and metrics have been developed for the task of image classification, our experiments are limited to this application. To fully demonstrate the generalisability of `MetaQuantus`, we plan to extend our experiments more broadly in the sciences and medicine and to include other data domains such as tabular, textual and time series data in the future. This will require additional work to ensure that the metrics themselves support these data domains, which will be addressed in upcoming publications.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9525–9536, 2018.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7786–7795, 2018.

Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Gimenez-Abalos. Focus! rating xai methods and finding biases. *CoRR*, abs/2203.02928, 2021. doi: 10.48550/arXiv.2109.15035.

Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Inf. Fusion*, 81:14–40, 2022.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, 2010.

Naman Bansal, Chirag Agarwal, and Anh Nguyen. 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.

I. Bellido and E. Fiesler. Do backpropagation trained neural networks have normal weight distributions? In Stan Gielen and Bert Kappen (eds.), *ICANN '93*, pp. 772–775, London, 1993. Springer London. ISBN 978-1-4471-2063-6.

José Manuel Benítez, Juan Luis Castro, and Ignacio Requena. Are artificial neural networks black boxes? *IEEE Trans. Neural Networks*, 8(5):1156–1164, 1997. doi: 10.1109/72.623216. URL https://doi.org/10.1109/72.623216.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 3016–3022. ijcai.org, 2020.

Alexander Binder, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations, 2022. URL https://arxiv.org/abs/2211.12486.

Lennart Brocki and Neo Christopher Chung. Evaluation of interpretability methods and perturbation artifacts in deep neural networks. *CoRR*, abs/2203.02928, 2022. doi: 10.48550/arXiv.2203.02928. URL https://doi.org/10.48550/arXiv.2203.02928.

Lukas Brunke, Prateek Agrawal, and Nikhil George. Evaluating input perturbation methods for interpreting CNNs and saliency map comparison. In *Computer Vision – ECCV 2020 Workshops*, pp. 120–134. Springer International Publishing, 2020. doi: 10.1007/978-3-030-66415-2_8.

Kirill Bykov, Anna Hedström, Shinichi Nakajima, and Marina M.-C. Höhne. Noisegrad: enhancing explanations by introducing stochasticity to model weights. *CoRR*, abs/2106.10185, 2021. URL https://arxiv.org/abs/2106.10185.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1383–1391. PMLR, 2020.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. *CoRR*, abs/2202.00734, 2022. URL https://arxiv.org/abs/2202.00734.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13567–13578, 2019.

Arne Gevaert, Axel-Jan Rousseau, Thijs Becker, Dirk Valkenborg, Tijl De Bie, and Yvan Saeys. Evaluating feature attribution methods in the image domain. *CoRR*, abs/2202.12270, 2022. URL https://arxiv.org/abs/2202.12270.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations. *CoRR*, abs/2202.06861, 2022.

Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Narine Kokhlikyan, Vivek Miglani, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. Investigating sanity checks for saliency maps with image and text classification. *arXiv preprint arXiv:2106.07475*, 2021.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *CoRR*, abs/2202.01602, 2022. URL https://arxiv.org/abs/2202.01602.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Simon Letzgus, Patrick Wagner, Jonas Lederer, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. Toward explainable AI for regression models. *CoRR*, abs/2112.11407, 2021. URL https://arxiv.org/abs/2112.11407.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4765–4774, 2017.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15, 2018.

Niels J. S. Mørch, Ulrik Kjems, Lars Kai Hansen, Claus Svarer, Ian Law, Benny Lautrup, Stephen C. Strother, and Kelly Rehm. Visualization of neural networks using saliency maps. In *Proceedings of International Conference on Neural Networks (ICNN'95), Perth, WA, Australia, November 27 - December 1, 1995*, pp. 2085–2090. IEEE, 1995.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. Order in the court: Explainable AI methods prone to disagreement. *CoRR*, abs/2105.03287, 2021. URL https://arxiv.org/abs/2105.03287.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *CoRR*, abs/2007.07584, 2020. URL https://arxiv.org/abs/2007.07584.

Frederik Pahde, Galip Ümit Yolcu, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Optimizing explanations by network canonization and hyperparameter search, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Laura Rieger and Lars Kai Hansen. IROF: a low resource evaluation metric for explanation methods. *CoRR*, abs/2003.08747, 2020. URL https://arxiv.org/abs/2003.08747.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18770–18795. PMLR, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Networks Learn. Syst.*, 28(11):2660–2673, 2017.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016. URL http://arxiv.org/abs/1605.01713.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified BP attributions fail. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9046–9057. PMLR, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mukund Sundararajan and Ankur Taly. A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values. *CoRR*, abs/1806.04205, 2018. URL http://arxiv.org/abs/1806.04205.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pp. 1474–1484. IEEE, 2022.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun D. Preece. 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.

Yipei Wang and Xiaoqian Wang. A unified study of machine learning explanation evaluation metrics. *CoRR*, abs/2203.14265, 2022. doi: 10.48550/arXiv.2203.14265. URL https://doi.org/10.48550/arXiv.2203.14265.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Mengjiao Yang and Been Kim. Benchmarking Attribution Methods with Relative Feature Importance. *CoRR*, abs/1907.09701, 2019.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10965–10976, 2019.

Gal Yona and Daniel Greenfeld. Revisiting sanity checks for saliency maps. *CoRR*, abs/2110.14297, 2021. URL https://arxiv.org/abs/2110.14297.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.*, 126(10):1084–1102, 2018.

A Appendix

In this section, we include all the necessary details and information to support the claims and results presented in the main paper. In Section A.1, we present theoretical considerations for the meta-evaluation framework. In Sections A.2 and A.3, we outline the mathematical definitions of explanation quality, for categories and estimators, respectively. In Section A.4, we describe the experimental setup and in the following Section A.5, we describe the experiments that were performed to sanity-check the framework. In Section A.6, we provide supplementary results, in terms of additional applications and supporting experiments. We provide a notation table at the end in Section A.7.

Broader Impact Statement

This research is important since we raise awareness and address the need for more reliable evaluation methods in the Explainable AI (XAI) community. In the XAI community, the evaluation of explanation methods has often been neglected or clouded by the ambiguity that an absence of ground truth labels entails—yet to foster sustainable progress in the field over time it is necessary to systematically define and evaluate the methods used to measure the quality of explanations. This research takes the first step towards this goal by developing practical, quantifiable tools for reliable evaluation. Without careful examination of the quality of explanations, the deployment of potentially beneficial machine learning algorithms may be hindered, preventing the full potential of AI from being realised in important areas such as healthcare, education, finance and policy.

A.1 Theoretical Considerations

This section discusses the concept of minor and disruptive perturbations in the context of multi-label classification and regression tasks in XAI. It also addresses the potential vulnerability of the framework to adversarial attacks and the motivation for the calculation of the IEC scoring criterion for disruptive perturbations.

A.1.1 Minor and Disruptive Perturbations

Multi-label classification In this work, given the popularity of image classification tasks in XAI, we focused mostly on this application. In this application, the definitions of minor and disruptive perturbations (as given in Definitions 1 and 2) apply to $\hat{y}$ and y' which represent the predicted classes for true and perturbed samples, respectively. However, our definitions can be easily extended to other types of classification tasks, such as multi-label classification. In this case, for a multi-label classification problem with C classes, the definitions of minor and disruptive perturbations (as given in Definitions 1 and 2) would apply to binary prediction vectors $\hat{\boldsymbol{y}} \in \mathbb{R}^C$ and $\boldsymbol{y}' \in \mathbb{R}^C$, rather than single classes. The distance between the two vectors can be denoted using the $|||\cdot|||_n$ notation.

Regression As is the case with many explanation methods (Letzgus et al., 2021), the extension to the regression problem in XAI is not straightforward. Given y and y' as real-valued prediction outcomes, we would need to adjust Definitions 1 and 2 to encompass a derivation of a proper boundary ϵ. We leave the task of adapting the meta-evaluation framework to regression problems to future work.

Adversarial attacks Adversarial attacks are techniques used to manipulate or deceive models (Szegedy et al., 2014) or their explanations (Dombrowski et al., 2019) by introducing perturbations to the input data that are imperceptible to humans but results in an incorrect prediction by the model. To adversarially attack Definition 1, it is theoretically possible to define a perturbation that maximises the strength of the perturbation, while still remaining consistent with Definition 1, i.e., $||\hat{y} - y'||_p > \epsilon$. In the same vein, to attack Definition 2, it is theoretically possible to define a perturbation that minimises the strength of the perturbation, while still remaining consistent with Definition 2, i.e., $||\hat{y} - y'||_p \leq \epsilon$. While it may be a theoretical possibility to attack the framework through the definitions, performing such attacks would not serve any practical purpose. This is because it contradicts the purpose of the framework, which is to assist practitioners in selecting and developing reliable XAI methods.

A.1.2 Motivation for IEC_{AR} calculation

Contrary to the calculation of the inter-consistency criterion for minor perturbations, i.e., IEC_{NR}, we cannot motivate IEC_{AR} based on ranking consistency with respect to explanation methods, as disruptive perturbations implicate a change in the estimator score. While the expectation of changed rankings as a result of disruptive perturbations may appear intuitive, the imposed change on the quality estimators could theoretically lead to a symmetrical change across explanation method scores, which preserves the ranking across explanation methods. Since the behaviour of the explanation functions under disruptive perturbations lies in the unverifiable spaces $\mathbb{U}$, we cannot exclude the possibility of a symmetrical response. Accordingly, for the calculation of IEC_{AR}, we relax the theoretical assumptions to a ranking comparison based on scores (as defined in Equation 7) which remain in the verifiable spaces Ω.

The assumption for the calculation of the IEC score with respect to disruptive perturbation is motivated by the scenario of an ideal estimator, which is expected to be able to assess the true performance of an explanation method, denoted q^{true}. In the ideal scenario, the real performance varies only slightly, i.e., $q_j^{\text{true}} \pm \epsilon$ would therefore define an upper estimation bound $q_j^{\text{true}} \approx q_j^{\max}$ for each explanation method $j \in [1, \dots, L]$[3]. All estimates $\bar{Q}_{i,j}^D$ resulting from the AR scenario should differ from the unperturbed quality estimate $\bar{Q}_{i,j}^D \neq \bar{Q}_{i,j}^*$. In the idealised scenario $q_j^{\text{true}} \approx q_j^{\max}$, we argue that $\bar{Q}_{i,j}^* \approx q_j^{\text{true}}$ and $\bar{Q}_{i,j}^D < \bar{Q}_{i,j}^*$, leading to Equation 8. Note, however, that in practice quality estimates are subject to larger variations which means that the assumption $q_j^{\text{true}} \approx q_j^{\max}$ and Equation 8 might not always hold. Therefore, in practice, we do not expect $\text{IEC}_{AR} \approx 1$, which aligns with our results in Table 1. Nonetheless, further research on the inter-consistency criterion under disruptive perturbations is subject to future work.

A.2 Explanation Quality: Category Definitions

In the main paper, we described how a lack of explanation ground truth labels has led to a diverse set of interpretations of explanation quality. In the following, we provide a brief summary of the most established categories of explanation quality, grouped into six categories; (a) faithfulness, (b) robustness, (c) localisation, (d) randomisation, (e) complexity and (f) axiomatic metrics (Hedström et al., 2022). To establish a mathematical ground for each category, we present a summarising equation. This means that all the nuances that typically exist within a category of explanation quality is not considered. For completeness, we, therefore, provide the exact mathematical descriptions of each of the individual estimators used in this work in Appendix A.3.

Since many explanation categories do rely on perturbation, we define a general perturbation function on any real-valued space $\mathbb{S} \subseteq \mathbb{R}^N, N \in \mathbb{N}$ in the following.

Definition 3 (Perturbation)**.** Let $\mathcal{P}_{\mathbb{S}}(\boldsymbol{s}; \eta) : \mathbb{S} \mapsto \mathbb{S}$ be a perturbation function of $\boldsymbol{s} \in \mathbb{S}$ with parameters $\eta \in \mathbb{R}$ such that $\forall \hat{\boldsymbol{s}} \in \mathbb{S}, \hat{\boldsymbol{s}} \neq \boldsymbol{s}$:

$$\mathcal{P}_{\mathbb{S}}(\boldsymbol{s}; \eta) = \hat{\boldsymbol{s}}. \tag{10}$$

For simplicity, we also write $\mathcal{P}_{\mathbb{S}}(\boldsymbol{s}) =: \mathcal{P}_{\mathbb{S}}(\boldsymbol{s}; \eta)$, which is used in the main paper.

Faithfulness (Montavon et al., 2018; Samek et al., 2017; Bach et al., 2015; Bhatt et al., 2020; Nguyen & Martínez, 2020) quantifies the extent that explanations follow the predictive behaviour of the model, asserting that more important features affect model decisions more strongly. Given f, $\boldsymbol{x}$, y' and $\hat{\boldsymbol{e}}$, the change in the model output $f(\boldsymbol{x})$ is measured as the input features of $\boldsymbol{x}$ are manipulated based on their attribution in $\hat{\boldsymbol{e}}$. The input manipulation is defined as a perturbation function $\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}, M)$ with $\boldsymbol{x} \in \mathbb{X}$ where M is the number of input features that are perturbed. Since f denotes a trained model with parameters θ, for brevity, we denote $f(\boldsymbol{x}; \theta)$ as $f(\boldsymbol{x})$ where possible.

$$\Psi_{\text{F}}(\Phi, f, \boldsymbol{x}, M) = |(f(x) - f(\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}, M))|. \tag{11}$$

Robustness (Montavon et al., 2018; Alvarez-Melis & Jaakkola, 2018; Yeh et al., 2019; Dasgupta et al., 2022) measures the stability of the explanation function with respect to small changes in the input, requiring

[3]Here, we present general theoretical considerations, but the specific claims for each metric would require individual proofs.

that those small perturbations in the input space $||\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}) - \boldsymbol{x}||_p < \varepsilon$, e.g., under an ℓ_p norm constraint upper bounded by some positive constant ε, lead to only slight changes in the explanation $||\hat{\boldsymbol{e}} - \Phi(\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}), f, \hat{y})|| < \varepsilon$ assuming that the model output approximately stayed the same $f(\boldsymbol{x}) \approx f(\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}))$.

$$\Psi_{\text{RO}}(\Phi, f, \boldsymbol{x}, \hat{y}, \mathcal{P}) = ||\hat{\boldsymbol{e}} - \Phi(\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}), f, \hat{y}; \lambda)|| \leq \varepsilon. \tag{12}$$

Localisation (Theiner et al., 2022; Kohlbrenner et al., 2020; Zhang et al., 2018; Rong et al., 2022; Arias-Duart et al., 2021) tests if the explainable evidence is centred around a region of interest, which may be defined around an object by a bounding box, a segmentation mask or a cell within a grid. It requires an additional segmentation mask $\boldsymbol{s}^{gt} \in \mathbb{R}^D$, mostly a binary mask of the input $\boldsymbol{s}_i^{gt} \in \{0, 1\}$, serving as a simulation or "proxy" of ground truth. While many variations exist, the goodness of $\hat{\boldsymbol{e}}$ can be defined by, e.g., their intersection divided by their union.

$$\Psi_{\text{L}}(\hat{\boldsymbol{e}}, \boldsymbol{s}^{gt}) = \frac{\hat{\boldsymbol{e}} \cap \boldsymbol{s}^{gt}}{\hat{\boldsymbol{e}} \cup \boldsymbol{s}^{gt}}. \tag{13}$$

Randomisation (Adebayo et al., 2018; Sixt et al., 2020) measures the extent explanations deteriorate as randomness is introduced to the evaluation. For example, Adebayo et al. (2018) measure the change in explanation as model parameters θ are increasingly randomised, requiring large perturbations in the parameter space of the model, i.e., $\mathcal{P}_{\mathbb{F}}(\theta) \gg \varepsilon$ to result in large changes in the explanation, i.e., $||\hat{\boldsymbol{e}} - \Phi(\boldsymbol{x}, \mathcal{P}_{\mathbb{F}}(\theta), \hat{y}; \lambda)|| \gg \varepsilon$.

$$\Psi_{\text{RA}}(\Phi, f, \boldsymbol{x}, \hat{y}, \varepsilon) = ||\hat{\boldsymbol{e}} - \Phi(\boldsymbol{x}, \mathcal{P}_{\mathbb{F}}(\theta), \hat{y}; \lambda)|| \gg \varepsilon. \tag{14}$$

Complexity (Chalasani et al., 2020; Bhatt et al., 2020; Nguyen & Martínez, 2020) captures the conciseness of explanations, i.e., only a few features should be selected to explain a model prediction. The notion of complexity differs in how it is empirically interpreted, e.g., by computing the Shannon entropy of attribution map (Bhatt et al., 2020). Alternatively, Chalasani et al. (2020) quantifies complexity by calculating the Gini Index of the absolute value of the attribution vector $\hat{\boldsymbol{e}}$ where D is the length of the attribution vector.

$$\Psi_{\text{C}}(\hat{\boldsymbol{e}}) = \frac{\sum_{i=1}^{D}(2i - D - 1)\hat{\boldsymbol{e}}_i}{D \sum_{i=1}^{D} \hat{\boldsymbol{e}}_i}. \tag{15}$$

Axiomatic (Sundararajan et al., 2017; Nguyen & Martínez, 2020) metrics gauges to what extent an explanation fulfil some axiomatic properties such as completeness (Sundararajan et al., 2017) and non—sensitivity (Nguyen & Martínez, 2020). Due to the ambiguity that arises when empirically evaluating metrics in this category, we do not study this category in detail.

A.3 Explanation Quality: Estimator Definitions

Within each of the five categories of explanation quality used in this work; (a) faithfulness, (b) robustness, (c) localisation, (d) randomisation and (e) complexity, we selected two estimators per category in our experiments.

Faithfulness Correlation (FC) (Bhatt et al., 2020) is defined in the following:

$$\Psi_{\text{FC}} = \underset{S \in |S| \subseteq d}{\text{corr}} \left(\sum_{i \in S} \Phi(\boldsymbol{x}, f, \hat{y}; \lambda)_i, f(\boldsymbol{x}) - f\left(\boldsymbol{x}_{[\boldsymbol{x}_s = \overline{\boldsymbol{x}}_s]}\right) \right), \tag{16}$$

where $|S| \subseteq D$ is a subset of indices of a sample $\boldsymbol{x}$, $\overline{\boldsymbol{x}}$ is the chosen baseline value and $\boldsymbol{x}_{[\boldsymbol{x}_s = \overline{\boldsymbol{x}}_s]}$ is, therefore, the masked input, with indices chosen randomly. Since f denotes a trained model with parameters θ, for brevity, we denote $f(\boldsymbol{x}; \theta)$ as $f(\boldsymbol{x})$ where possible. Higher values indicate that the explanation method's assignment of attribution is correlated with the behaviour of the model and hence is preferred.

Pixel-Flipping (PF) (Bach et al., 2015) returns a curve of prediction scores over an iterative set of pixel replacements, which are sorted in descending order by the highest relevant pixel in the explanation $\Phi(\boldsymbol{x}, f, \hat{y}; \lambda)$. To return one evaluation score per input sample, we calculate the area under the curve (AUC) as follows:

$$\Psi_{\text{PF}} = \sum_{i=1}^{n} (\hat{y}_i + \hat{y}_{i+1}) \cdot \frac{p_{i+1} - p_i}{2} \tag{17}$$

where n is the number of discrete perturbation steps, p_i and p_{i+1} are the x-values for the ith and $(i+1)^{th}$ perturbation steps and $\hat{y}_i$ and $\hat{y}_{i+1}$ are the prediction values. For faithful explanations, a steep degradation of prediction scores is expected when attributions are iteratively replaced in descending order. Therefore, a lower value of AUC is indicative of better performance.

Max-Sensitivity (MS) (Yeh et al., 2019) measures the maximum sensitivity of an explanation using a Monte Carlo sampling-based approximation. It is defined as follows:

$$\Psi_{\text{MC}} = \max_{\boldsymbol{x}+\delta \in \mathcal{N}_\epsilon(\boldsymbol{x}) \leq \varepsilon} \left[\frac{\|\Phi(\boldsymbol{x}, f, \hat{y}; \lambda) - \Phi(\boldsymbol{x}+\delta, f, \hat{y}; \lambda)\|}{\|\boldsymbol{x}\|} \right], \tag{18}$$

where ε defines the radius of a discrete, finite-sample neighborhood around each input sample $\boldsymbol{x}$. This neighborhood, denoted as $\mathcal{N}_\epsilon(\boldsymbol{x})$, includes all samples in the set X that are within a distance of ε from $\boldsymbol{x}$. A lower MS score is indicative of more robustness.

Local Lipschitz Estimate (LLE) (Alvarez-Melis & Jaakkola, 2018) works similarly to the Max-Sensitivity (MS) method and estimates the Lipschitz constant of the explanation, which is a measure of how much the explanation changes with respect to the input under slight perturbation, defined as δ. The LLE method is defined as follows:

$$\Psi_{\text{LLE}} = \max_{\boldsymbol{x}+\delta \in \mathcal{N}_\epsilon(\boldsymbol{x}) \leq \epsilon} \frac{\|\Phi(\boldsymbol{x}, f, \hat{y}; \lambda) - \Phi(\boldsymbol{x}+\delta, f, \hat{y}; \lambda)\|_2}{\|\boldsymbol{x} - (\boldsymbol{x}+\delta)\|_2}, \tag{19}$$

where lower values indicate less change with respect to the change in input, which is desirable.

Pointing-Game (PG) (Zhang et al., 2018) captures whether the feature of maximal attribution lies on the ground truth mask, which is a binary mask indicating the true features that contribute to the model's output. It is defined as follows:

$$\Psi_{\text{PG}} = \begin{cases} 1 & \text{if } \arg\max_i \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda) \in \boldsymbol{s}^{gt} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where $\Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)$ represents the ith input feature of highest atttribution and $\boldsymbol{s}^{gt} \in \mathbb{R}^D$ denotes the binary ground truth mask.

Relevance Mass Accuracy (RMA) (Arras et al., 2022) quantifies the fraction of the sum of the attribution that intersects with the ground truth mask over the full explanation sum. It is defined as follows:

$$\Psi_{\text{RMA}} = \frac{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda) \cdot \boldsymbol{s}_{gt,i}}{\sum_{i=1}^{D} \Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)}, \tag{21}$$

where $\Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)$ is the attribution of the ith input feature and $\boldsymbol{s}_{gt,i}$ is the value of the ith element in the ground truth mask.

Model Parameter Randomisation Test (MPR) (Adebayo et al., 2018) measures the correlation between an explanation from a randomly parameterised model $f(\boldsymbol{x}; \mathcal{P}_{\mathbb{F}}(\theta; v)) = \hat{f}$ and the original model f for

each separate layer v of the network. To generate one quality estimate per sample, we calculate the average of the correlation coefficients over all the layers in the network, denoted V:

$$\Psi_{\text{MPR}} = \frac{1}{V} \sum_{v=1}^{V} \text{corr}(\Phi^v(\boldsymbol{x}, f, \hat{y}; \lambda), \Phi^v(\boldsymbol{x}, \hat{f}, \hat{y}; \lambda)), \tag{22}$$

where $\Phi^v(\boldsymbol{x}, f, \hat{y}; \lambda)$ is the explanation generated by the original model f for layer v and $\Phi^v(\boldsymbol{x}, \hat{f}, \hat{y}; \lambda)$ is the explanation generated by the randomly parameterised model $\hat{f}$ for layer v. The correlation between the two explanations is calculated for each layer and then averaged over all layers to generate the MPR score, where a lower correlation coefficient is desired.

Random Logit (RL) method proposed by (Sixt et al., 2020) is originally defined using the structural similarity index ($SSIM$) over the explanation of the ground truth label and an explanation of non-target class y'. However, to make it comparable with the MPR metric, the *SSIM* calculation is replaced with the *Spearman Rank Correlation Coefficient* as follows:

$$\Psi_{\text{RL}} = \text{corr}(\Phi(\boldsymbol{x}, f, \hat{y}; \lambda), \Phi(\boldsymbol{x}, f, y'; \lambda)), \tag{23}$$

where $\Phi(\boldsymbol{x}, f, \hat{y}; \lambda)$ is the explanation generated for the prediction $\hat{y}$ and $\Phi(\boldsymbol{x}, f, y'; \lambda)$ is the explanation generated for a non-target class y'. Lower values indicate that the explanations are not correlated which is desirable.

Sparseness (SP) (Chalasani et al., 2020) is a method for evaluating the sparsity of explanations and is defined as the Gini index of the explanation. It is calculated by summing the product of the ranks of the input features and their attributions and dividing by the sum of the attribution as follows:

$$\Psi_{\text{SP}} = \frac{\sum_{i=1}^{D} (2i - D - 1) \cdot \hat{\boldsymbol{e}}_i}{D(D-1) \sum_{i=1}^{D} \hat{\boldsymbol{e}}_i}, \tag{24}$$

a higher sparseness score indicates lower complexity of the explanation $\hat{\boldsymbol{e}}$, which is desirable.

Complexity (CO) (Bhatt et al., 2020) is defined using the Shannon entropy calculation which measures the amount of uncertainty or randomness in the explanation map. It is calculated by summing the product of the probabilities of the attributions and the logarithm of the probabilities of the attributions:

$$\begin{aligned} \Psi_{\text{CO}} = \mathbb{E}_i \left[-\ln \left(\mathbb{P}_\Phi \right) \right] = -\sum_{i=1}^{D} \mathbb{P}_\Phi(i) \ln \left(\mathbb{P}_\Phi(i) \right) \\ \text{with} \quad \mathbb{P}_\Phi(i) = \frac{|\Phi_i(\boldsymbol{x}, f, \hat{y}; \lambda)|}{\sum_{j \in [d]} |\Phi_j(\boldsymbol{x}, f, \hat{y}; \lambda)|}; \mathbb{P}_\Phi = \{\mathbb{P}_\Phi(1), \ldots, \mathbb{P}_\Phi(d)\}, \end{aligned} \tag{25}$$

where $|\cdot|$ denotes the absolute value, $\mathbb{P}_\Phi(i)$ denotes the fractional contribution of feature $\boldsymbol{x}_i$ to the total quantity of the attribution. A higher entropy indicates a higher level of uncertainty or randomness, i.e., a higher complexity. A uniformly distributed attribution would have the highest possible complexity score.

A.4 Experimental Setup

In this section, we describe the experimental setup more in detail, which includes the datasets, models, explanation methods and estimators used in this work. We keep this section short since most of the methods in the following have been widely discussed in previous works. For more details, we refer the reader to the respective original publications. The experiments can be reproduced following the instructions in the repository (`anonymous`).

Datasets We use four image classification datasets in the experiments— MNIST(LeCun et al., 2010), fMNIST (Xiao et al., 2017), cMINST (customised-MNIST) (Bykov et al., 2021) and ILSVRC-15 (ImageNet) (Russakovsky et al., 2015). For MNIST and fMNIST, we randomly sample 1024 test samples. We also randomly sampled 384 test samples from cMINST (customised-MNIST) (Bykov et al., 2021) which consists of MNIST digits displayed on uniformly sampled CIFAR-10 (Krizhevsky et al., 2009) backgrounds. To understand the real impact of State-of-the-art (SOTA), we also perform our some experiments on ILSVRC-15 (ImageNet) (Russakovsky et al., 2015), using 206 randomly selected test samples.

We have chosen these datasets based on the availability of segmentation masks, since the estimators within the localisation category require such masks for computation. The bounding boxes for these datasets are designed to enclose the object of interest. For cMNIST, the bounding box covers 25% of the input and for MNIST and fMNIST, they cover approximately 35% (but up to 64%) of the input features. For ImageNet, the bounding boxes vary in size depending on the class of interest.

Models The experiments are performed using different neural network models, including architectures such as LeNets (LeCun et al., 1998) and ResNets (He et al., 2016) which contributes to the robustness of our findings. For MNIST and fMNIST, we train LeNets to an accuracy of 98.14% and 87.44% respectively. For the cMNIST dataset, a ResNet-9 is trained to an accuracy of 98.09%. The training of all models is performed in a similar fashion; employing SGD optimisation with a standard cross-entropy loss, an initial learning rate of 0.001 and momentum of 0.9. All models are trained for 20 epochs each. For ILSVRC-15 (Russakovsky et al., 2015), we use the ResNet-18 model with pre-trained weights given the ImageNet dataset, accessible via `PyTorch` (Paszke et al., 2019).

Explanation methods We employ explanation methods from a widely used category of post-hoc attribution methods, both gradient-based and model-agnostic techniques, i.e., *Gradient* (Mørch et al., 1995; Baehrens et al., 2010), *Saliency* (Mørch et al., 1995), *GradCAM* (Selvaraju et al., 2020), *Integrated Gradients* (Sundararajan et al., 2017), *Input×Gradient* (Shrikumar et al., 2016), *Occlusion* (Zeiler & Fergus, 2014) and *GradientSHAP* (Lundberg & Lee, 2017).

In all experiments, we generate explanations for a sample's predicted class $\hat{y}$. Whereas certain estimators such as the Saliency explanation ignore the signs of the explanations, we refrain from taking their absolute values, to preserve the explainable evidence in the attribution. For comparability, we normalise the explanations prior to the evaluation analysis using the square root of its average second-moment estimate (Binder et al., 2022), which is defined as follows:

$$\frac{\hat{\boldsymbol{e}}_{h,w}}{\left(\frac{1}{HW}\sum_{h',w'}\hat{\boldsymbol{e}}^2_{h',w'}\right)^{1/2}}\,, \tag{26}$$

where $\hat{\boldsymbol{e}}_{h,w}$ is the value of the explanation map at pixel location $(h,\, w)$ and H, W denote the height and width, respectively[4].

Estimators We select the most established metrics within each of the five categories of explanation quality: *Complexity* (CO) (Bhatt et al., 2020), *Sparseness* (SP) (Chalasani et al., 2020), *Faithfulness Correlation* (FC) (Bhatt et al., 2020), *Pixel-Flipping* (PF) (Bach et al., 2015), *Max-Sensitivity* (MS) (Yeh et al., 2019), *Local Lipschitz Estimate* (LLE) (Alvarez-Melis & Jaakkola, 2018), *Pointing-Game* (PG) (Zhang et al., 2018), *Relevance Mass Accuracy* (RMA) (Arras et al., 2022), *Model Parameter Randomisation Test* (MPR) (Adebayo et al., 2018) and *Random Logit* (RL) (Sixt et al., 2020). We have defined each of the individual metrics mathematically in Appendix A.3.

Parameterisation For the initialisation of the different estimators, we mostly followed the recommendations as stated in the respective original publications. However, to make the metrics within a certain explanation category as comparable as possible, alterations to certain hyperparameters were made. When

[4]This normalisation ensures that each score in the attribution map has an average squared distance to zero that is equal to one. Since this operation does not normalise the attributions into a fixed range, it is not meant for visualisations, rather it is meant to preserve a quantity that is useful for the comparison of distances between different explanation methods.

applying *Pixel-Flipping* (Bach et al., 2015) on image datasets, it generally becomes computationally infeasible to iterate over one pixel at a time. Therefore, we iterate over $\frac{2*w}{D}$ where D is the dimensions of the input and w and h are the width and height of the image, respectively (which are assumed to be the same). We also use this same value to choose the subset size for *Faithfulness Correlation* (Bhatt et al., 2020). For both faithfulness metrics, as the replacement strategy for masked pixels, we use uniform sampling where we set the lower and higher bounds to the minimum and the maximum value of the test set, respectively. For the robustness metrics, which both are based on Monte Carlo sampling-based approximation, we let it run for 10 iterations. In the randomisation category, for comparability, we use the *Spearman's Rank Correlation Coefficient* to calculate the similarity between the original explanation and the explanation subject to randomisation. A full overview of the parameterisation of the metrics can be found in the GitHub repository `anonymous`.

Hardware All experiments were computed on GPUs where we used NVIDIA A100-PCIE 40GB for the toy datasets and NVIDIA A100-PCIE 80GB for ImageNet dataset.

A.5 Sanity-Checks of the Meta-Evaluation Framework

In this section, we conduct two sanity-checking experiments. In the first experiment, we create and meta-evaluate adversarial estimators to demonstrate the usability of the framework in practice and highlight how the two failure modes act complementary. In the second experiment, we examine the extent that the choice of L, i.e., the set of explanation methods, may influence the MC score.

A.5.1 Adversarial Estimators

To sanity-check the meta-evaluation framework, we created adversarial quality estimators that were intended to perform poorly in a specific failure mode and thus, should indisputably fail the corresponding part of the testing scenarios of IPT and MPT. Specifically, we created an adversarial quality estimator that, independent of its given model, data and explanations, returns scores that are always the same (i.e., using deterministic sampling[5]). As such, this estimator should ultimately fail the reactivity to adversary tests (i.e., IAC_{AR} and IEC_{AR}) since those tests expect a response to disruption. We denote this estimator $\Psi_{=}$. We create a second adversarial quality estimator that, independent of its inputs, returns scores that are drawn from a different probability distribution (i.e., using stochastic sampling[6]). In this setup, we expect poor performance on the noise resilience tests (i.e., IAC_{NR} and IEC_{NR}) since these tests check that the quality estimates remain relatively unchanged after perturbation. We denote this adversarial estimator $\Psi_{\neq}$.

Table 2 summarises the outcome of this exercise, which includes the four score elements IAC_{NR}, IAC_{AR}, IEC_{NR} and IEC_{AR}, aggregated for 5 iterations with $K = 10$ for the two tests, IPT and MPT. The expectation of the test outcome is indicated by the value in brackets after the display of the actual score, including the standard deviation. Here, a value of 0 indicates that the test should fail[7] and any other value indicates the desired outcome of the test to be successful. From Table 2, we note that both estimators, $\Psi_{\neq}$ and $\Psi_{=}$ produced scores that align with the expected value. Since estimator $\Psi_{\neq}$, relies on stochastic sampling, the scores are approximate, nevertheless, the scores and expectation are close and the standard deviation is small, indicating that the sanity checks results are stable. Overall, we can observe that the expectation aligns with the empirical reality across the different test settings. Therefore, we conclude the sanity-checking experiment to be passed.

Another important insight that can be drawn from Table 2 is that the two failure modes complement each other in determining the performance of an estimator. For an estimator that is provably bad, i.e., returns

[5]We assemble this adversarial estimator by repeatedly returning the same scores for q' as one set of uniformly sampled scores $\hat{q} \sim \mathcal{U}(\alpha, \beta)$ with $\alpha = 0$ and $\beta = 1$.

[6]Here, we sample from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 = 1$ but with different means for the unperturbed- and the perturbed estimates, respectively. For the unperturbed estimates $\hat{q}$, we sample μ from a wide range, i.e., $\mu \in [-100000, -1]$ and for the perturbed estimates q', we set a narrow range with $\mu \in [0, 1]$.

[7]The exception is the expected value of the inter-consistency score, IEC_{NR}, for estimator $\Psi\neq$ is not 0.0 but 0.25. This is because, for an estimator that assigns attributions randomly, i.e., independent of the explanation method, the likelihood of the condition $\bar{r}_j^M = \bar{r}_j^\star$ is $\frac{1}{L}$.

scores that are completely unrelated to the model, data and explanation methods (such as demonstrated by $\Psi\neq$ and $\Psi=$), at least one of the failure modes (AR or NR) will reveal that the estimator is failing. To fully assess the performance of an estimator, both failure modes are therefore necessary.

Table 2: The sanity-check exercise results show aggregated scores including std, over 5 iterations with $K = 5$. The direction of the arrow, i.e., ↑ indicates if a higher value is better. The expectation of the test outcome is indicated by the value in brackets, after the display of the actual score.

Test	**Estimator**	$\mathbf{IAC}_{NR}$ (↑)	$\mathbf{IAC}_{AR}$ (↑)	$\mathbf{IEC}_{NR}$ (↑)	$\mathbf{IEC}_{AR}$ (↑)
IPT	$\Psi_{\neq}$	0.015 ± 0.023 (0.00)	0.983 ± 0.011 (1.00)	0.248 ± 0.004 (0.25)	0.0 ± 0.0 (0.00)
	$\Psi_{=}$	1.0 ± 0.0 (1.00)	0.0 ± 0.0 (0.00)	1.0 ± 0.0 (1.00)	0.0 ± 0.0 (0.00)
MPT	$\Psi_{\neq}$	0.014 ± 0.010 (0.00)	0.973 ± 0.019 (1.00)	0.248 ± 0.003 (0.25)	0.0 ± 0.0 (0.00)
	$\Psi_{=}$	1.0 ± 0.0 (1.00)	0.0 ± 0.0 (0.00)	1.0 ± 0.0 (1.00)	0.0 ± 0.0 (0.00)

A.5.2 Dependency on L

The meta-evaluation framework is intentionally designed to take into account the set of explanation methods given in the setup. For example, in the inter-consistency criterion (IEC) for noise resilience, we compute the estimator's ability to rank different explanation methods consistently when exposed to minor perturbations. The resulting MC score of a quality estimator will, therefore, to a certain extent, be dependent on the choice of L: both in terms of its cardinality and how similar the explanation functions are.

To understand how the performance of our quality estimator may vary depending on the choice of L, we conducted an experiment where we computed the MC score while enumerating various choices of L. In this experiment, we vary both the cardinality of L, by choosing values of $\{2, 3, 4\}$ and the methods included in the set. We selected both model-agnostic explanation methods such as *Occlusion* (Zeiler & Fergus, 2014) as well as gradient-based techniques such as *GradCAM* (Selvaraju et al., 2020), *Integrated Gradients* (Sundararajan et al., 2017) etc. For the sets of 2 explanation methods we included: {*Gradient, Occlusion*}, {*Gradient, Input×Gradient*}, {*Gradient, Saliency*}, {*Gradient, Input×Gradient*}. For sets with 3 methods: {*Gradient, GradCAM, GradientSHAP*}, {*Gradient, Saliency, Integrated Gradients*} and for 4 methods: {*Gradient, Saliency, Input×Gradient, GradCAM*}, {*Gradient, Saliency, Occlusion, GradCAM*}.

In Figure 7, we show the aggregate values for different explanation sets across the datasets separately. Here, the error bars indicate the standard deviations. By comparing the MC scores category by category, we can observe that the error bars from the respective metric do generally not overlap. This means that the choice of L has limited influence on the MC score, suggesting the measure's stability.

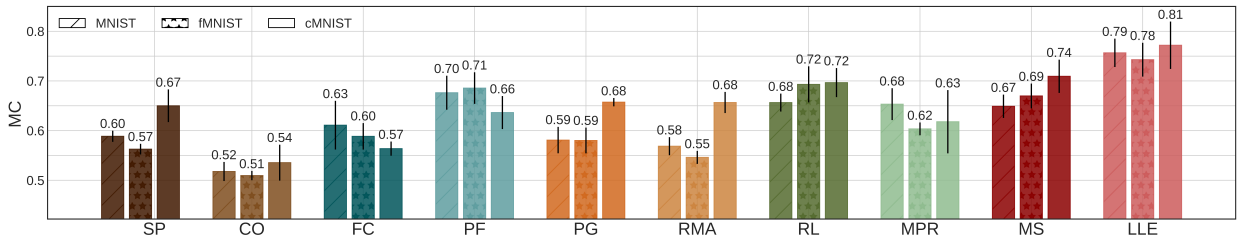


Figure 7: Comparison of averaged meta-consistency performance for different quality estimators using MPT and IPT, aggregated over 3 iterations with $K = 5$, across models {*LeNet, ResNet*} and different datasets {*MNIST, fMNIST, cMNIST*} with error bars showing the standard deviation.

A.6 Supplementary Experiments

In the following section, we present additional experiments conducted in the scope of this work. First, we demonstrate that `MetaQuantus` can be used for additional applications in Explainable AI. Here, we include two demonstrations, first, we show how the MC score can be employed as a target variable for optimising the hyperparameters of a metric and second, we demonstrate how the framework can be used to analyse

category convergence. At the end of this section, we discuss supplementary results for the benchmarking experiment which includes an additional analysis of ranking consistency.

A.6.1 Application — Hyperparameter Optimisation

It is practically well-known and increasingly publicly recognised (Bansal et al., 2020; Brunke et al., 2020; Brocki & Chung, 2022; Pahde et al., 2022) how difficult it can be to tune the hyperparameters in the explainability domain. Unlike in traditional machine learning, in XAI, we generally do not have a target variable to optimise against. As an additional experiment, we, therefore, investigated how the meta-evaluation framework can be useful in solving the task of selecting the best set of hyperparameters for a given metric. For this, we choose a metric with relatively many parameters, that is *Faithfulness Correlation* (Bhatt et al., 2020) and performed a grid-search on these using ImageNet. By exploring combinations of three baseline strategies = ['Black', 'Uniform', 'Mean'] and four subset sizes = $[28, 52, 102, 128]$, we created 12 hyperparameter settings[8]. We determined the performance of each of the metric's parameterisation by storing the meta-evaluation vector $\boldsymbol{m}$ and the MC score at each run. The objective of this exercise is to determine the hyperparameter setting that optimises the performance of the estimator, i.e., its resilience to noise and reactivity to adversary.

From Figure 8 (left), we can observe that P11 has the highest meta-consistency score and as such, we recommend the associated parameter setting of "mean" as the replacement strategy with 102 features as the subset size. In contrast to previous works that found a relatively large difference in evaluation outcomes between different parameterisations of faithfulness metrics (Tomsett et al., 2020; Brunke et al., 2020; Rong et al., 2022; Hedström et al., 2022), we detect, that with the MC score—which provides a more comprehensive picture of the estimator's performance—there is not a considerable variability, as depicted by the similarity in IAC and IEC scores over P1 to P12.

A.6.2 Application — Category Convergence

The question of whether evaluation metrics within the same category are measuring the same underlying concept has been of significant interest to the community (Tomsett et al., 2020; Gevaert et al., 2022). Based on the observed similarity of estimator shapes in Figure 5—that the estimators within the same category typically have a higher resemblance in area shapes compared to estimators outside of their categories—we sought to employ the meta-evaluation framework to investigate whether metrics within a category exhibit

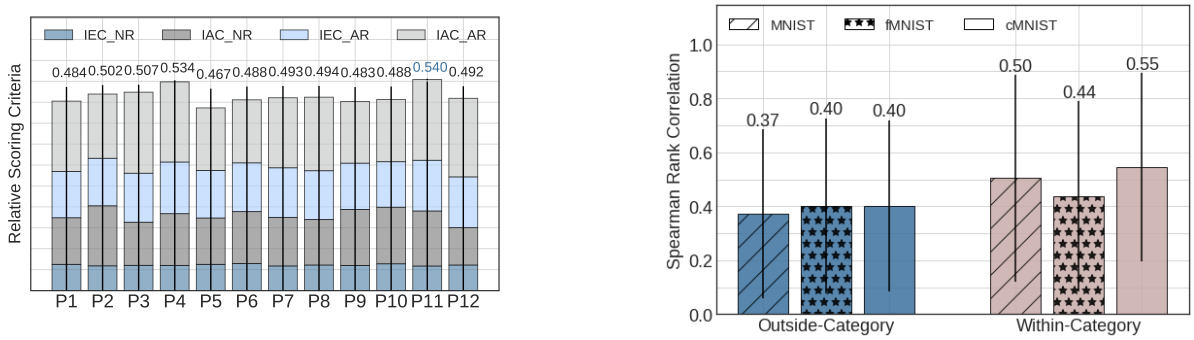


Figure 8: *Left:* The results of using the meta-evaluation framework to optimize the hyperparameters of FC (Bhatt et al., 2020) metric across 12 parameterisations (P1-P12) on ImageNet dataset, averaged over 3 iterations with $K = 3$. The parameter setting P11 demonstrated the highest scores with small standard deviation and thus is selected as the parameter setting. *Right:* The results from comparing the correlation coefficients between the meta-evaluation vector scores for estimators within the same category versus those outside of the category, suggesting that the estimators of the same category have more resemble with respect to its performance characteristics compared to estimators outside.

[8]The parameters were combined in the following 12 settings: P1: ['Black', 28], P2: ['Black', 52], P3: ['Black', 102], P4: ['Black', 128], P5: ['Uniform', 28], P6: ['Uniform', 52], P7: ['Uniform', 102], P8: ['Uniform', 128], P9: ['Mean', 28], P10: ['Mean', 52], P11: ['Mean', 102], P12: ['Mean', 128].

a greater level of correlation than those outside of the category. To address this question, often referred to as "convergent validity", the prevalent technique has been to measure intra-correlation, which simply involves correlating the scores of different estimators within the same category. This approach, however, has limitations, as it disregards the aspect of ranking consistency (IEC) and may not account for the fact that scores from different estimators may have vastly different scales and interpretations, which may skew the results.

We improve upon the current methodology proposed in (Tomsett et al., 2020; Gevaert et al., 2022) by calculating the correlation coefficient on the meta-evaluation vector $\boldsymbol{m}$ of different estimators, within- and outside of their category as produced in the benchmarking exercise. This approach is advantageous as it: (i) yields scores in a normalised range $[0, 1]$ and (ii) provides a more comprehensive view of the estimator's performance characteristics by incorporating multiple failure modes and criteria.

Figure 8 (right) presents the results of this experiment, averaged over all estimators as described in A.4. Here, we can observe that the estimator's performance characteristics are more similar within a category, as seen in the higher correlation coefficient (*Spearman Rank Correlation Coefficient*) across all datasets. These observations contrast previous works by (Tomsett et al., 2020; Gevaert et al., 2022) that found a low correlation coefficient (for faithfulness estimators in particular). We posit that this difference can be explained by the fact that the meta-evaluation framework considers multiple failure modes and criteria of what a quality estimator should fulfil and not only one, e.g., ranking consistency (Rong et al., 2022) and as such, give a more comprehensive answer. However, from the error bars in Figure 8, we also observe, that the correlation coefficients are greatly varying within each group. Further research is thus necessary to fully understand the extent to which estimators of the same explanation quality category measure the same underlying concept.

A.6.3 Supplementary Results — Benchmarking

In the following, similar to Table 1, we present the results of the Input Perturbation Test and the Model Perturbation Test for the fMNIST and cMNIST datasets in Tables 3 and 4, respectively. Tables 3 and 4 can be found at the end of this Section. The grey rows indicate the results from the Input Perturbation Test and the white rows show the results from the Model Perturbation Test. The results are consistent with those presented in the main manuscript, both in terms of individual score criteria and top-performing estimators in each category.

Similar to Figure 6, we also represent Tables 3 and 4 as area graphs. With an exception of slightly higher localisation scores for cMNIST dataset (as explained in the main paper), the results as demonstrated in Figures 9-10 are completely consistent with those findings presented in the main paper. Recall that, larger coloured areas imply better performance on the different scoring criteria and the grey area indicates the area of an optimally performing quality estimator, i.e., $\boldsymbol{m}^* = \mathbb{1}^4$. Each column of estimators represents a category of explanation quality, from left to right: *Complexity*, *Faithfulness*, *Localisation*, *Randomisation* and *Robustness*.

Similar to Figure 5, we also visualise the results (as shown in Table 3-4) as scatter plots for fMNIST and cMNIST datasets in Figure 11. In the main paper, we identified that the faithfulness category (blue points) had particularly lower ranking consistency (IEC), which is also evident in these supplementary plots. From Figure 11, we can moreover observe how the estimators' scores on the respective failure modes are related. Figure 11 shows that a higher resilience to noise does not necessarily imply more reactivity to adversary and vice versa—the performance characteristics of the estimators are more complex than that.

A.6.4 Supplementary Results — Ranking Consistency

In the main paper, we presented the average MC scores for each dataset in Figure 5, which showed consistency across tested datasets, with the best-performing estimator in each category of explanation quality remaining consistent across datasets. To further explore this consistency, we considered a margin of error of 2 standard deviations applied to the MC scores and re-calculated the within-category ranking for each individual estimator in each category and visualised the results in Figure 12.

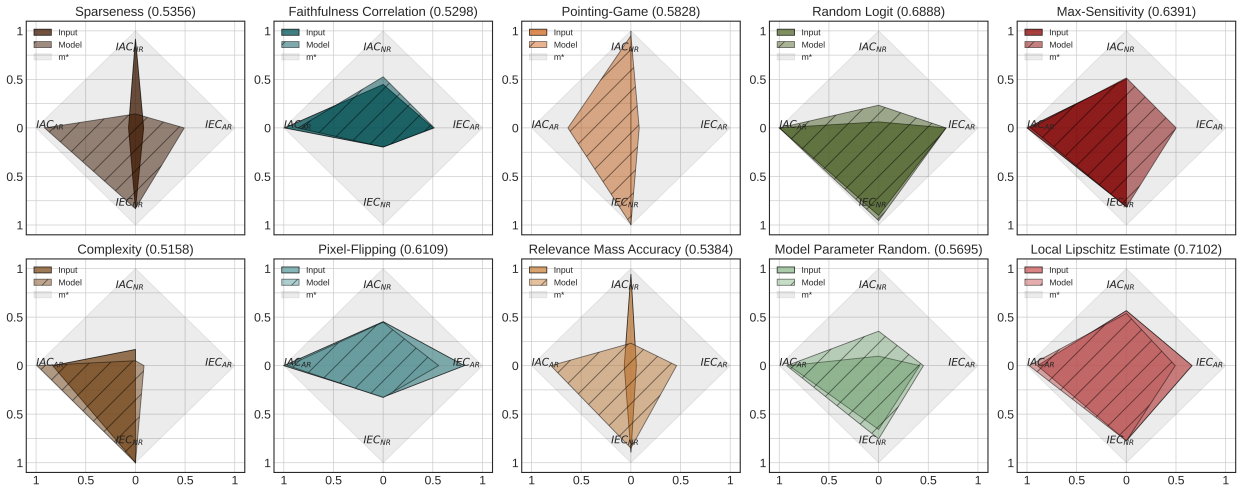


Figure 9: A graphical representation of the benchmarking results (Table 3), aggregated over 3 iterations with $K = 5$. Each column corresponds to a category of explanation quality, from left to right: *Complexity*, *Faithfulness*, *Localisation*, *Randomisation* and *Robustness*. The grey area indicates the area of an optimally performing estimator, i.e., $\mathbf{m}^* = \mathbb{1}^4$. The MC score (indicated in brackets) is averaged over MPT and IPT. Higher values are preferred.

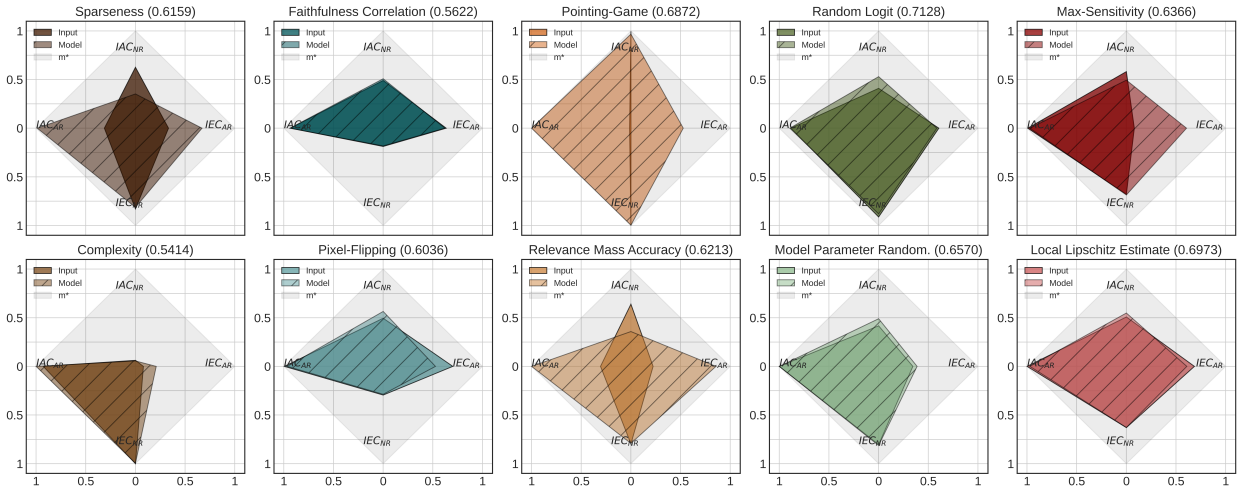


Figure 10: A graphical representation of the benchmarking results (Table 4), aggregated over 3 iterations with $K = 5$. Each column corresponds to a category of explanation quality, from left to right: *Complexity*, *Faithfulness*, *Localisation*, *Randomisation* and *Robustness*. The grey area indicates the area of an optimally performing estimator, i.e., $\mathbf{m}^* = \mathbb{1}^4$. The MC score (indicated in brackets) is averaged over MPT and IPT. Higher values are preferred.

Figure 12 showcase the distribution of the frequency with which the different estimators within each category were ranked as the highest or the lowest, respectively. The colour scheme used is in line with previous figures, where larger fractions indicate more frequent "wins". From Figure 12, we infer that for MPT there are few instances where the best-performing estimator is ranked second, indicating stability in the results. On the other hand, for IPT, the difference between the best- and worst-performing estimator is smaller, where we often observe that the rankings are reversed. Since the MC scores reported in the main paper are computed by averaging over both MPT and IPT, a variability in rankings is possible. It is important for practitioners of `MetaQauantus` to be aware of the possible variability in rankings, by means of exercising caution when drawing conclusions about the relative performance of individual metrics.

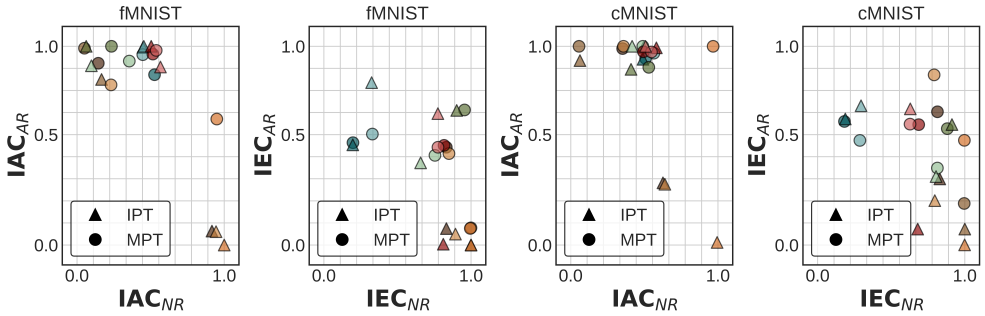
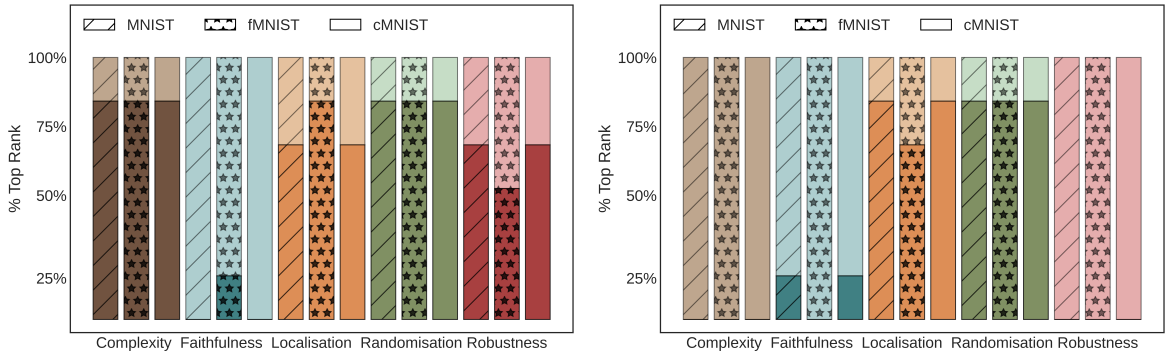


Figure 11: A supplementary visualisation of the benchmarking results (Table 3-4), in particular IAC and IEC scores for noise resilience (x-axes) and adverse reactivity (y-axes). The colours indicate the estimator and the symbols demonstrate the test: IPT and MPT, respectively. Higher values are preferred.



(a) Model Perturbation Test

(b) Input Perturbation Test

Figure 12: A supplementary visualisation of the benchmarking results (Tables 1, 3 and 4) showing the distribution of top rankings within each category of explanation quality. For the MPT tests there is little variability in rankings, but for IPT test it is higher.

Table 3: Benchmarking results for fMNIST dataset, aggregated over 3 iterations with $K = 5$. IPT results are in grey rows and MPT results are in white rows. The symbol $\overline{\text{MC}}$ denotes the averages of the MC scores over IPT and MPT. The top-performing method in each category, which outperforms the bottom-performing method by at least 2 standard deviations, is underlined. Higher values are preferred for all scoring criteria.

Category	**Estimator**	$\overline{\textbf{MC}}$ (↑)	**MC** (↑)	**IAC**$_{NR}$ (↑)	**IAC**$_{AR}$ (↑)	**IEC**$_{NR}$ (↑)	**IEC**$_{AR}$ (↑)
Complexity	Sparseness	0.536 ± 0.011	0.596 ± 0.012	0.145 ± 0.039	0.915 ± 0.045	0.831 ± 0.004	0.492 ± 0.082
			0.475 ± 0.010	0.917 ± 0.036	0.070 ± 0.003	0.832 ± 0.003	0.083 ± 0.001
	Complexity	0.516 ± 0.007	0.532 ± 0.014	0.050 ± 0.047	0.990 ± 0.027	0.999 ± 0.000	0.086 ± 0.028
			0.500 ± 0.000	0.167 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
Faithfulness	Faithfulness Corr.	0.530 ± 0.021	0.524 ± 0.021	0.527 ± 0.030	0.857 ± 0.072	0.198 ± 0.008	0.515 ± 0.004
			0.536 ± 0.021	0.448 ± 0.087	0.994 ± 0.003	0.196 ± 0.004	0.504 ± 0.002
	Pixel-Flipping	0.530 ± 0.021	0.573 ± 0.025	0.447 ± 0.050	0.958 ± 0.088	0.329 ± 0.002	0.558 ± 0.032
			0.649 ± 0.018	0.453 ± 0.073	1.000 ± 0.000	0.324 ± 0.001	0.817 ± 0.003
Localisation	Pointing-Game	0.583 ± 0.005	0.666 ± 0.009	0.950 ± 0.025	0.634 ± 0.032	0.995 ± 0.001	0.084 ± 0.018
			0.500 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
	Relevance Rank Acc.	0.538 ± 0.023	0.587 ± 0.024	0.231 ± 0.102	0.806 ± 0.056	0.850 ± 0.007	0.460 ± 0.048
			0.490 ± 0.022	0.944 ± 0.034	0.067 ± 0.067	0.894 ± 0.003	0.055 ± 0.003
Randomisation	Random Logit	0.689 ± 0.005	0.717 ± 0.010	0.234 ± 0.039	1.000 ± 0.000	0.955 ± 0.005	0.680 ± 0.005
			0.660 ± 0.000	0.062 ± 0.000	1.000 ± 0.000	0.902 ± 0.000	0.677 ± 0.000
	Model Param. Rand.	0.570 ± 0.010	0.622 ± 0.010	0.355 ± 0.042	0.925 ± 0.000	0.755 ± 0.005	0.451 ± 0.000
			0.518 ± 0.010	0.098 ± 0.008	0.902 ± 0.045	0.657 ± 0.004	0.414 ± 0.001
Robustness	Max-Sensitivity	0.639 ± 0.036	0.699 ± 0.037	0.515 ± 0.097	0.961 ± 0.021	0.816 ± 0.007	0.501 ± 0.058
			0.580 ± 0.035	0.504 ± 0.141	1.000 ± 0.000	0.811 ± 0.002	0.004 ± 0.000
	Local Lipschitz Est.	0.710 ± 0.022	0.696 ± 0.038	0.538 ± 0.139	0.979 ± 0.033	0.775 ± 0.005	0.492 ± 0.092
			0.724 ± 0.006	0.567 ± 0.037	0.896 ± 0.024	0.774 ± 0.001	0.661 ± 0.006

Table 4: Benchmarking results for cMNIST dataset, aggregated over 3 iterations with $K = 5$. IPT results are in grey rows and MPT results are in white rows. The symbol $\overline{\text{MC}}$ denotes the averages of the MC scores over IPT and MPT. The top-performing method in each category, which outperforms the bottom-performing method by at least 2 standard deviations, is underlined. Higher values are preferred for all scoring criteria.

Category	**Estimator**	$\overline{\textbf{MC}}$ (↑)	**MC** (↑)	**IAC**$_{NR}$ (↑)	**IAC**$_{AR}$ (↑)	**IEC**$_{NR}$ (↑)	**IEC**$_{AR}$ (↑)
Complexity	Sparseness	0.616 ± 0.015	0.706 ± 0.013	0.352 ± 0.061	0.989 ± 0.017	0.814 ± 0.001	0.670 ± 0.016
			0.525 ± 0.018	0.626 ± 0.099	0.313 ± 0.028	0.830 ± 0.005	0.333 ± 0.006
	Complexity	0.541 ± 0.018	0.565 ± 0.024	0.056 ± 0.084	1.000 ± 0.000	0.996 ± 0.001	0.209 ± 0.013
			0.518 ± 0.013	0.062 ± 0.010	0.928 ± 0.047	1.000 ± 0.000	0.080 ± 0.005
Faithfulness	Faithfulness Corr.	0.562 ± 0.014	0.563 ± 0.017	0.508 ± 0.061	0.939 ± 0.017	0.182 ± 0.004	0.622 ± 0.005
			0.562 ± 0.010	0.490 ± 0.031	0.934 ± 0.018	0.188 ± 0.008	0.634 ± 0.012
	Pixel-Flipping	0.604 ± 0.016	0.586 ± 0.022	0.565 ± 0.040	0.965 ± 0.022	0.287 ± 0.005	0.526 ± 0.080
			0.621 ± 0.010	0.495 ± 0.037	0.995 ± 0.001	0.295 ± 0.012	0.701 ± 0.002
Localisation	Pointing-Game	0.687 ± 0.006	0.873 ± 0.010	0.967 ± 0.000	1.000 ± 0.000	0.997 ± 0.000	0.527 ± 0.040
			0.502 ± 0.001	0.995 ± 0.003	0.013 ± 0.003	0.999 ± 0.001	0.001 ± 0.000
	Relevance Rank Acc.	0.621 ± 0.011	0.856 ± 0.020	0.751 ± 0.008	0.358 ± 0.055	1.000 ± 0.000	0.791 ± 0.012
			0.491 ± 0.014	0.640 ± 0.028	0.306 ± 0.032	0.796 ± 0.003	0.223 ± 0.005
Randomisation	Random Logit	0.713 ± 0.005	0.723 ± 0.010	0.530 ± 0.065	0.894 ± 0.026	0.881 ± 0.007	0.586 ± 0.012
			0.703 ± 0.000	0.410 ± 0.000	0.884 ± 0.000	0.913 ± 0.000	0.606 ± 0.000
	Model Param. Rand.	0.657 ± 0.009	0.673 ± 0.003	0.490 ± 0.006	1.000 ± 0.000	0.814 ± 0.005	0.387 ± 0.000
			0.641 ± 0.016	0.417 ± 0.058	1.000 ± 0.000	0.804 ± 0.005	0.344 ± 0.004
Robustness	Max-Sensitivity	0.637 ± 0.030	0.690 ± 0.035	0.494 ± 0.069	0.972 ± 0.045	0.687 ± 0.004	0.606 ± 0.050
			0.583 ± 0.024	0.582 ± 0.079	0.992 ± 0.002	0.680 ± 0.019	0.080 ± 0.002
	Local Lipschitz Est.	0.697 ± 0.020	0.689 ± 0.026	0.548 ± 0.077	0.971 ± 0.049	0.628 ± 0.007	0.609 ± 0.042
			0.706 ± 0.014	0.508 ± 0.047	0.999 ± 0.000	0.630 ± 0.007	0.685 ± 0.005

A.7 Notation Table

Preliminaries

f	A black-box model function that maps input $\boldsymbol{x}$ to output y
θ	The parameters of the model function f
$\boldsymbol{X}_{\text{tr}}$	The training dataset on which the model f is trained
$\boldsymbol{X}_{\text{te}}$	The test dataset on which the model f is evaluated
$\boldsymbol{x}$	An input in the instance space $\mathbb{X}$
y	An output class in the label space $\mathbb{Y}$
$\hat{y}$	A prediction made by the model f
C	The number of output classes
N	The number of test samples
D	The dimension of the input
$\mathbb{X}$	The instance space
$\mathbb{Y}$	The label space
$\mathbb{F}$	The function space of all models
Φ	An explanation function that maps $\boldsymbol{x}$, f, and $\hat{y}$ to an explanation map $\hat{\boldsymbol{e}}$
λ	The parameter of the explanation function Φ
$\hat{\boldsymbol{e}}$	The explanation map produced by Φ
$\mathbb{E}$	The space of possible explanations
Ψ	A quality estimation function that takes $\hat{\boldsymbol{e}}$ and returns a scalar $\hat{q}$ to indicate its quality
τ	The parameter of the quality estimation function Ψ
$\hat{q}$	A quality estimate made by the estimator Ψ
$\mathbb{O}$	The verifiable spaces of the estimator's Ψ input parameters $\mathbb{O} \in \{\{\mathbb{X}\}, \{\mathbb{F}\}, \{\mathbb{X}, \mathbb{F}\}\}$
$\mathbb{U}$	The unverifiable spaces of the estimator's Ψ input parameters $\mathbb{U} \in \{\{\mathbb{E}\}, \{\mathbb{O}\}, \{\mathbb{E}, \mathbb{O}\}\}$

A Meta-Evaluation Framework

NR	The first failure mode, noise resilience
AR	The second failure mode, adversary reactivity
y'	A prediction after perturbation on the input, model or both input and model spaces
ϵ	A threshold $\epsilon \in \mathbb{R}$ for determining the type of perturbation
t	The perturbation strength $t \in \{M, D\}$
$\mathcal{P}_{\mathbb{\Omega}}(\boldsymbol{\omega})$	A perturbation function of the verifiable spaces $\boldsymbol{\omega} \in \mathbb{\Omega}$
$\mathcal{P}^M_{\mathbb{X}}$	A minor perturbation function of the input space $\mathbb{X}$
$\mathcal{P}^M_{\mathbb{F}}$	A minor perturbation function of the function space $\mathbb{F}$
$\mathcal{P}^D_{\mathbb{X}}$	A disruptive perturbation function of the input space $\mathbb{X}$
$\mathcal{P}^D_{\mathbb{F}}$	A disruptive perturbation function of the function space $\mathbb{F}$
K	The number of perturbations
L	The set of explanation methods
$\hat{\boldsymbol{q}}$	The unperturbed quality estimates $\hat{\boldsymbol{q}} \in \mathbb{R}^N$
$\boldsymbol{q}'_k$	The perturbed quality estimates, replicated K times for N test samples
d	A statistical significance function that takes $\hat{\boldsymbol{q}}$ and $\boldsymbol{q}'_k$ and returns a p-value
r	A ranking function that takes nominal values and returns rankings in descending order
$\boldsymbol{Q}$	A matrix of all perturbed samples over K perturbations
$\bar{\boldsymbol{Q}}$	A matrix for the unperturbed estimates $\hat{\boldsymbol{q}}$ for L explanation methods, averaged over K
$\bar{\boldsymbol{Q}}'$	A matrix for the perturbed estimates $\boldsymbol{q}'_k$ for L explanation methods, averaged over K
$\bar{\boldsymbol{Q}}^M$	A matrix for the perturbed estimates under minor perturbation
$\bar{\boldsymbol{Q}}^D$	A matrix for the perturbed estimates under disruptive perturbation
$\boldsymbol{U}$	A binary ranking agreement matrix that takes quality estimates from $\bar{\boldsymbol{Q}}$ and $\bar{\boldsymbol{Q}}'$ and populates the entries according to the interpretation of ranking
$\boldsymbol{U}^M$	A binary ranking agreement matrix with perturbed estimates under minor perturbation
$\boldsymbol{U}^D$	A binary ranking agreement matrix with perturbed estimates under disruptive perturbation
$\boldsymbol{m}$	A meta-consistency vector containing the IAC and IAC scores for both failure modes
$\boldsymbol{m}^*$	An optimally performing quality estimator Ψ as defined by the all-one vector $\mathbb{1}^4$
IAC	The intra-consistency scoring criterion, where IAC $\in [0, 1]$
IEC	The inter-consistency scoring criterion, where IEC $\in [0, 1]$
MC	The meta-consistency score, where MC $\in [0, 1]$

Practical Evaluation

$\mathcal{U}$	The uniform distribution with parameters α, β
α	The lower bound of the uniform distribution $\mathcal{U}(\alpha, \beta)$
β	The upper bound of the uniform distribution $\mathcal{U}(\alpha, \beta)$
$\boldsymbol{\delta}_i$	Additive uniform noise applied to input space such that $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \boldsymbol{\delta}_i$
$\mathcal{N}$	The normal distribution with parameters $\boldsymbol{\mu}, \Sigma$
$\boldsymbol{\mu}$	The mean of the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$
Σ	The variance of the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$
$\boldsymbol{\nu}_i$	Multiplicative Gaussian noise applied applied to model parameters such that $\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta} \cdot \boldsymbol{\nu}_i$