Evaluate with the Inverse: Efficient Approximation of Latent Explanation Quality Distribution

Carlos Eiras-Franco[1]
Anna Hedström[2,3,5], **Marina M.-C. Höhne**[4,5]
[1]Universidade da Coruña. CITIC
[2]UMI Lab, Leibniz Institute of Agricultural Engineering and Bioeconomy e.V. (ATB)
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[4]Data Science Department, Leibniz Institute of Agricultural Engineering and Bioeconomy e.V. (ATB)
[5]Department of Computer Science, University of Potsdam
carlos.eiras.franco@udc.es

Abstract

Obtaining high-quality explanations of a model's output enables developers to identify and correct biases, align the system's behavior with human values, and ensure ethical compliance. Explainable Artificial Intelligence (XAI) practitioners rely on specific measures to gauge the quality of such explanations. These measures assess key attributes, such as how closely an explanation aligns with a model's decision process (faithfulness), how accurately it pinpoints the relevant input features (localization), and its consistency across different cases (robustness). Despite providing valuable information, these measures do not fully address a critical practitioner's concern: how does the quality of a given explanation compare to other potential explanations? Traditionally, the quality of an explanation has been assessed by comparing it to a randomly generated counterpart. This paper introduces an alternative: the Quality Gap Estimate (QGE). The QGE method offers a direct comparison to what can be viewed as the 'inverse' explanation, one that conceptually represents the antithesis of the original explanation. Our extensive testing across multiple model architectures, datasets, and established quality metrics demonstrates that the QGE method is superior to the traditional approach. Furthermore, we show that QGE enhances the statistical reliability of these quality assessments. This advance represents a significant step toward a more insightful evaluation of explanations that enables a more effective inspection of a model's behavior.

1 Introduction

Model output explanations play a crucial role in AI alignment by enhancing transparency, understandability, and trustworthiness in AI systems. In most contexts related to explainability, ground-truth explanations are unavailable (Dasgupta, Frost, and Moshkovitz 2022; Hedström et al. 2023). This absence inherently complicates the task of evaluating explanations. Consequently, efforts to evaluate explanations vary widely, ranging from assessing the robustness of explanations to noise, their complexity, and their localization, to evaluating how faithfully an explanation represents the underlying model.

Although it is not possible to develop a metric based on verified ground truth, a crucial insight is that the adequacy of an explanation can still be assessed by comparing it relative to other explanations with the use of quality measures. Commonly, XAI evaluation methods yield numerical fitness measures that indicate the degree to which an explanation adheres to a certain criterion. However, these numerical values do not provide insight into an explanation's standing relative to the spectrum of possible explanations. In this study, we introduce a method that quantifies the quality of attribution-based explanation methods by efficiently estimating how an explanation's quality compares to the rest of the potential explanations. Despite its apparent simplicity, our evaluations using various sanity checks demonstrate that this strategy enhances the reliability of quality metrics.

While other studies have employed different notions of randomness to compare or evaluate explanation methods (Bommer et al. 2023; Samek et al. 2017; Adebayo et al. 2018), these approaches are often more computationally intensive and yield less favorable results. The proposed method is independent of the specific dataset, model, task, and, importantly, of the quality metric used.

The main contributions of this paper include:

- The introduction of the Quality Gap Estimate (QGE), a novel evaluation strategy that renders existing quality metrics more informative by aiding in the determination of an explanation's quality relative to alternatives with a limited increase of computational requirements.
- A redefinition of the problem of assessing the superiority of an explanation over its alternatives as a sampling problem, which generalizes the conventional method of comparison with a random explanation.
- An assessment of the applicability of QGE across a wide array of established quality metrics, evaluating its impact on various dimensions including faithfulness, localization, and robustness.
- The presentation of experimental findings that demonstrate an enhancement in the statistical reliability of existing quality metrics through the application of QGE, providing XAI practitioners with more reliable interpretation tools.

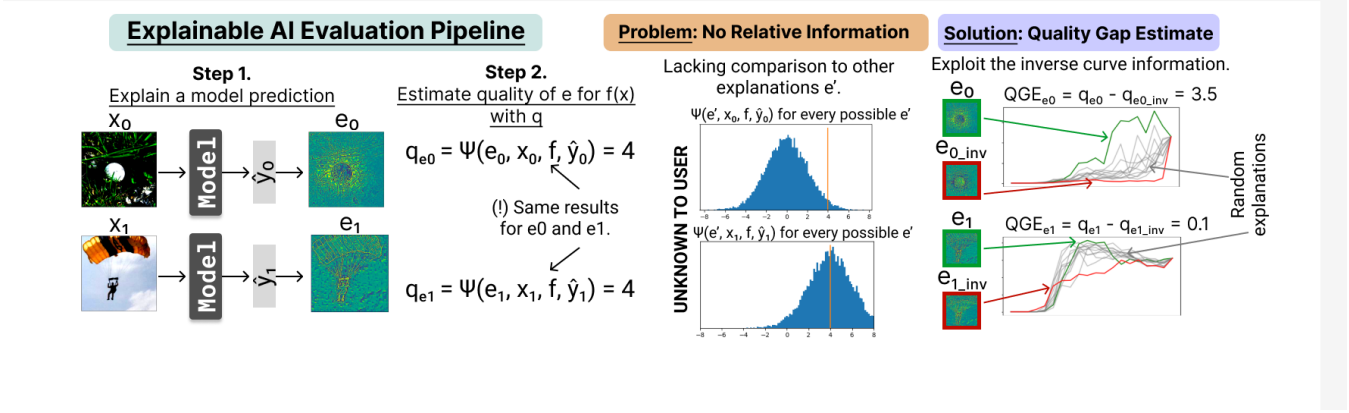


Figure 1: (Step 1). The usual XAI pipeline allows the user to obtain an explanation e for a prediction $\hat{y}$ using any explanation method. This is demonstrated for two distinct inputs ($\mathbf{x_0}$ and $\mathbf{x_1}$), producing predictions $\hat{y}_0$ and $\hat{y}1$, and explanations $\mathbf{e_0}$ and $\mathbf{e_1}$, respectively. (Step 2). To assess the quality of explanation $\mathbf{e}$ for prediction $\hat{y}$, the user computes a quality measure q. In this example, we use the area under the Pixel-Flipping curve, though the method can work with any attribution-based quality measure. (Problem) Despite both input/explanation pairs registering identical q values, it remains unknown to the user that $\mathbf{e_0}$ has higher quality than most explanations for the first prediction, while $\mathbf{e_1}$ has average quality compared to other explanations for the second prediction, as shown by their histograms. (Solution) To allow the user to effectively gauge the relative quality of explanation $\mathbf{e}$ against alternative explanations $\mathbf{e}'$, we introduce QGE, which measures the difference between the quality of $\mathbf{e}$ and the quality of $\mathbf{e}_{inv}$ (a rearrangement of $\mathbf{e}$ ranking features in reverse order). This comparative quality metric does not require costly sampling of the q distribution. Although both explanations have equivalent q values, using QGE, the user can discern that $\mathbf{e_0}$ is a high-quality explanation for the first prediction, while $\mathbf{e_1}$ is merely average for the second one. The user may then seek a better explanation for the second prediction.

2 Related Works

Evaluating the quality of methods without ground truth explanation labels is a significant challenge for researchers (Brunke, Agrawal, and George 2020; Brocki and Chung 2022; Rong et al. 2022; Hedström et al. 2023). Many studies have shown that faithfulness metrics are highly sensitive to their parameterisation during evaluation; altering patch sizes or pixel occlusion tactics can significantly affect the outcomes (Tomsett et al. 2020; Brunke, Agrawal, and George 2020; Brocki and Chung 2022; Rong et al. 2022; Hedström et al. 2023). These findings are concerning; if small changes in parameters cause large variations in evaluation outcomes, it may be hard to trust the results.

For more reliable estimations of explanation quality, individual explanation methods have been evaluated *relative* to a random baseline (Samek et al. 2017; Nguyen and Martinez 2020; Ancona et al. 2018). The concept of using the random explainer as a worst-case reference point has also been used for calculating explanation skill scores (Bommer et al. 2023) or as part of a paired t-test to compare with existing explanation methods (Rieger and Hansen 2020). Most similar to our work (Blücher, Vielhaben, and Strodthoff 2024) is incorporating information about the pixel-flipping inverse curve. Our contribution is different in both aim and applicability. Their approach aims at enhancing the occlusion process, specifically the masking strategy for pixel-flipping (Samek et al. 2017). We provide a general-purpose evaluative solution applicable across various explanation metrics such as localization, faithfulness, and robustness.

3 Method

3.1 A Framework for Explanation Quality

Consider a classification problem where we have a model, denoted as a function f that maps an input $\mathbf{x} \in \mathbb{R}^D$ to an output $\hat{y} \in \{1, \ldots, C\}$. This function estimates the probability distribution across classes i.e., $f(\mathbf{x})_i = p(y_i|\mathbf{x})\ \forall i \in \{1, \ldots, C\}$, so $\hat{y} = \text{argmax}(f(\mathbf{x}))$. To identify the features utilized by the model to predict $\hat{y}$, we employ an explanation function Φ as follows:

$$\Phi(\mathbf{x}, f, \hat{y}) = \mathbf{e} \tag{1}$$

Φ outputs a real-valued vector $\mathbf{e}$ with D components that assign attribution to each feature x_i in $\mathbf{x}$, indicating its relative importance in f's prediction of class $\hat{y}$.

Various methods exist to assess the suitability or fitness of $\mathbf{e}$ based on different attributes. Generally, we can define a quality measure Ψ as a function that evaluates the quality of a given explanation $\mathbf{e}$, relative to the model f, the input $\mathbf{x}$, and the predicted label $\hat{y}$. For brevity, we denote this scalar value as q_e, and we use simply q to refer to the function with a fixed f, $\mathbf{x}$, and $\hat{y}$ that evaluates a given explanation:

$$q_e := q(\mathbf{e}) := \Psi(\mathbf{e}, \mathbf{x}, f, \hat{y}) \tag{2}$$

3.2 The Need for Relative Information

The value q_e provides practitioners with a measure of how well an explanation $\mathbf{e}$ adheres to predefined quality criteria. It enables comparisons between different explanations by

contrasting q_e with $q_{e'}$. However, a crucial question remains for practitioners: How good is $\mathbf{e}$ compared to the whole set of alternative explanations? (see Figure 1).

To answer this question, we could check the position of q_e in the unknown distribution of q across all possible explanations $\mathbf{e}'$. Yet, this approach is impractical as it requires calculating the distribution of q for all possible explanations, which is computationally infeasible.

A strategy to tackle this problem consists of estimating the latent distribution of the quality measure by sampling q values. For instance, Pixel-Flipping (Bach et al. 2015) compares q_e against q_{e^r} which represents the quality score for a random explanation (denoted $\mathbf{e}^r$). This method effectively performs a single-sample estimation of the quality distribution. Pixel-Flipping transforms the original quality measure q into a new transformed measure:

$$qt(\mathbf{e}) = \Psi(\mathbf{e}, \mathbf{x}, f, \hat{y}) - \Psi(\mathbf{e^r}, \mathbf{x}, f, \hat{y}) = q_e - q_{e^r} \quad (3)$$

However, relying on just one random sample (q_{e^r}) can compromise the accuracy of the estimation. A more robust transformation would involve sampling multiple random explanations, computing their quality, and calculating an average. Yet, this method incurs higher computational costs.

To overcome these challenges, our goal is to develop a transformed quality measure qt that satisfies the following criteria:

(R1) Provides a value that clearly indicates the relative standing of q_e within the latent distribution of all possible $q_{e'}$ values. By evaluating $qt(\mathbf{e})$, a user should be able to determine whether $\mathbf{e}$'s quality is above-average, average, or below-average.

(R2) Preserves the comparative information inherent in q, especially its capacity to rank explanations. Given a explanation $\mathbf{e^i}$, any explanation $\mathbf{e^j}$ with higher quality, should also have a higher qt. More formally, qt should be constructed so that, given any pair of explanations (e^i, e^j), if $q_{e^i} < q_{e^j}$ holds, then $qt(\mathbf{e}^i) < qt(\mathbf{e}^j)$.

(R3) Is computationally efficient.

3.3 Proposed Method

We propose using the $\mathbf{e}^{inv}$ explanation, which ranks features in inverse order to $\mathbf{e}$, as an alternative to the commonly used random explanation $\mathbf{e}^r$. This approach aims to improve the quality of estimations while maintaining low computational cost. Given $o = \text{argsort}(\mathbf{e})$, we define

$$e^{inv}_{o_i} := e_{o_{D-i+1}} \quad \forall i \in [1..D] \quad (4)$$

where D is the number of variables in $\mathbf{e}$. Therefore, $\mathbf{e}^{inv}$ is a permutation of the values of $\mathbf{e}$ constructed so that the most attributed variable in $\mathbf{e}$ gets the smallest attribution in $\mathbf{e}^{inv}$, the second most attributed variable in $\mathbf{e}$ gets the second smallest attribution, and so on. $\mathbf{e}^{inv}$ is, then, the opposite interpretation of the original explanation $\mathbf{e}$. An example of this procedure is shown below.

$$\mathbf{e} = [0.1, -0.1, 9.0, 4.0] \quad o = \text{argsort}(\mathbf{e}) = [1, 0, 3, 2]$$
$$\mathbf{e}^{inv} = [4.0, 9.0, -0.1, 0.1] \quad \text{argsort}(\mathbf{e}^{inv}) = [2, 3, 0, 1]$$

By design, the ranking by attribution of the features in $\mathbf{e}^{inv}$ is the same as for $\mathbf{e}$, but in reverse order (i.e. argsort($\mathbf{e}$) = reversed(argsort($\mathbf{e}^{inv}$))).

Once we have $\mathbf{e^{inv}}$, we define the proposed transformation, which we will call its Quality Gap Estimation (QGE), as the difference between the quality value of the original explanation $\mathbf{e}$ and the quality value of $\mathbf{e^{inv}}$:

$$\texttt{QGE} = \Psi(\mathbf{e}, \mathbf{x}, f, \hat{y}) - \Psi(\mathbf{e^{inv}}, \mathbf{x}, f, \hat{y}) \quad (5)$$

The rationale behind this method is intuitive: QGE increases not only when $\Psi(\mathbf{e}, \mathbf{x}, f, \hat{y})$ is high, indicating the high quality of $\mathbf{e}$, but also when $\Psi(\mathbf{e}^{inv}, \mathbf{x}, f, \hat{y})$ is low, reflecting the poor quality of the inversely ranked explanation. A substantial gap between these values suggests that many random explanations would have quality values falling between these two, thereby indicating that $\mathbf{e}$ is of much higher quality than an average-quality explanation. Conversely, a small QGE implies that the quality difference between $\mathbf{e}$ and $\mathbf{e}^{inv}$ is minimal, suggesting that the order in which $\mathbf{e}$ ranks features is approximately as effective as any alternative, regardless of the absolute value of q_e. This pattern suggests that QGE satisfies requirement **R1**, with values approximately zero for average-quality explanations, negative for below-average, and positive for above-average explanations. Moreover, if **R2** is met, the more above-average $\mathbf{e}$'s quality is, the higher QGE will be, and similarly, the more below-average q_e's quality is, the lower QGE will be. The level of compliance with **R2** is assessed in Section 4.1.

Regarding **R3**, this requirement is met because QGE can be computed quickly. Generally, the cost of computing any transformation qt is dominated by the cost of computing Ψ, which is generally a costly function. The fewer times a transformation qt needs to compute Ψ, the faster it will be. Determining the QGE requires computing Ψ only twice (the original $\Psi(\mathbf{e}, \mathbf{x}, f, \hat{y})$, and the additional $\Psi(\mathbf{e^{inv}}, \mathbf{x}, f, \hat{y})$). This approach avoids the computational expense of needing multiple samples of Ψ for different explanations to estimate a distribution, as alternative methods do. Other than computing Ψ, QGE requires a single subtraction and computing $\mathbf{e}^{inv}$ as indicated in Eq. 4. While the time needed to compute $\mathbf{e}^{inv}$ is generally negligible compared to the cost of computing Ψ, this step can also be sped up in most cases. Although Eq. 4 preserves the original attribution values and reallocates them in reverse order, an alternate but straightforward computation could set $e^{inv}_i := -e_i \forall i \in [0..D]$ to achieve the goal of inversely ranking features compared to $\mathbf{e}$ (argsort(e) = reversed(argsort(e^{inv}))). However, the magnitudes of the attributions are not maintained after this transformation and, in instances where the quality metric Ψ requires bounded attribution values, users would need to adhere to the original formulation in Equation 4 if a shift and scale of $e^{inv}_i := -e_i$ is unsuitable.

An implementation of QGE for a wide variety of quality measures is available in the Quantus toolkit: https://github.com/understandable-machine-intelligence-lab/Quantus

4 Experimental Results

To verify the quality of our method we performed experiments focused on two main points:

- Assesing the level of compliance of QGE with the requirements listed in Section 3.2:

(R1) is met by the design of QGE (see intuitive explanation after Eq. 5). We report a complete exploration of the QGE distribution that confirms this.

(R2) Evaluating QGE in its ability to preserve the information inherent in the original quality metric q (**R2**). The details of this evaluation are discussed in Section 4.1.

(R3) The cost of computing a transformation qt is dominated by how many times it needs to compute the original quality function Ψ. Our experiments confirm this, and the results presented allow the comparison of QGE with an alternative of similar cost.

- Assessing the statistical reliability of QGE compared to competitive baseline methods. The experiments conducted for this are reported in Section 4.2.

In our experiments, we took a fixed model f, input sample $\mathbf{x}$, and label y. Although these parameters remained constant for each individual experiment, we varied them across different experiments to test the robustness and general applicability of our method.

The code used for all experiments is available at https://github.com/annahedstroem/eval-project

4.1 Suitability of QGE

For our initial experiments, we employed the Pixel-Flipping quality measure (Bach et al. 2015). Let $f_y(\mathbf{x})$ represent the output of model f for class y, and $\mathbb{P}(\mathbf{x}, \mathbf{e}, M)$ denote a perturbation function that modifies all features of $\mathbf{x}$ except for the M most attributed features according to explanation $\mathbf{e}$ (with $M \in [0, D]$). The quality of explanation $\mathbf{e}$ is then measured as the average value[1] of $f_y(\mathbf{x})$ over all possible levels of feature selection M:

$$q_e := \Psi(\mathbf{e}, \mathbf{x}, f, y) = \frac{1}{D} \sum_{m=0}^{D} f_y(\mathbb{P}(\mathbf{x}, \mathbf{e}, m)) \quad (6)$$

In this experimental setup, the perturbation function $\mathbb{P}(\mathbf{x}, \mathbf{e}, M)$ results in $\mathbf{x}'$, where the $D - M$ least relevant features (as determined by $\mathbf{e}$) are replaced with zeros. This introduces the well-known problem that $\mathbf{x}'$ is outside the distribution for which the model f was trained (Hase, Xie, and Bansal 2021), an issue that is addressed below (see "Advantage of using QGE vs. QRAND_1 across datasets and models").

We compared two different transformations: our evaluation measure, QGE, was compared against QRAND_K, a baseline measure that estimates the quality distribution across all explanations by sampling k random explanations. It calculates the difference between q_e and the average quality of those k samples, generalizing the usual comparison with a single random explanation.

$$\text{QRAND}_\text{K} = q_e - \frac{\sum_{i=0}^{K} q_{r_k}}{K} \quad (7)$$

[1]Some implementations measure the area under the curve instead of the average activation level. Both alternatives are equivalent since they are proportional.

where q_{r_k} represents the quality of a random explanation.

To assess adherence to **R2** for both quality metric transformations qt (either QGE or QRAND_K for a range of k values), we used the following evaluation metrics (both implemented in SciPy (Virtanen et al. 2020)):

- **Kendall rank correlation (τ):** As a direct translation of **R2**, we computed Kendall's rank correlation τ, i.e. the level of agreement of the order after the transformation with the order after the transformation (i.e. $q_i < q_j \implies qt_i < qt_j$ and $q_i > q_j \implies qt_i > qt_j$).
- **Spearman correlation (ρ)** between the transformed quality metric qt (either QGE or QRAND_K) and the original q.

Four scenarios are reported in Section 4.1: (a) the examination of the complete set of all possible explanations for two small datasets; (b) an exploration in larger datasets; (c) an analysis of the effect of the model used; and (d) the use of measures other than Pixel-Flipping.

a. Exhaustive exploration of the explanation space We first tested our method using two small datasets in which the complete list of all possible explanations can be enumerated in a reasonable time. It's crucial to note that differences in q_e and $q_{e'}$ arise solely when $\mathbf{e}$ and $\mathbf{e}'$ differ in their feature ranking orders. This constraint significantly simplifies the space of possible explanations and facilitates the enumeration process.

The datasets selected for this experiment were the Avila dataset (Stefano et al. 2018) and the Glass Identification dataset (German 1987), both small tabular datasets. Given the limited number of variables in these datasets, the complete set of distinct explanations (with respect to q; i.e. every attribution vector that yields a different order when argsorted) corresponds to the set of all possible permutations of the variables. This set can be exhaustively explored. The examples in Avila consist of 10 variables, which yields a total of $10! = 3,628,800$ different explanations while Glass has 9 variables, amounting to $9! = 362,880$ explanations.

Our analysis confirms that the distribution of QGE is centered on zero (see Appendix A.1). Consequently, if **R2** is met, (i.e. high-quality explanations obtain higher QGE values), then **R1** is also met since the user can clearly distinguish between explanations with an above-average quality (which have positive QGE) and explanations with below-average quality (which have negative QGE). The following experiments aim to assess the degree to which **R2** is met.

For each dataset, we trained a Multilayer Perceptron (MLP). The models achieved test accuracies of 0.99 on the Avila dataset and 0.77 on the Glass dataset, respectively. For each explanation $\mathbf{e}$ in the set of all possible explanations, we calculated q_e (Eq.6), QGE (Eq. 4), and QRAND_K (Eq. 7) for values of k ranging from 1 to 10.

Order preservation: Kendall's τ To measure the extent to which a transformation qt (either QRAND_K or QGE) preserves the order of the original quality measure q, we compute Kendall's $\tau_{q,qt}$.

In Fig. 2 we report the results for 5 different $\mathbf{x}$ inputs, for each of which all possible explanations were computed. These show that for `Avila`, on average, using `QGE` as a transformation maintains the correct ordering for 85% of pairs, which results in a $\tau_{q,\texttt{QGE}}$ of 0.7 ($\pm$ 0.12). To obtain a comparable result with the conventional $\texttt{QRAND}_\texttt{K}$, more than $k = 6$ random samples are needed. Similarly, for the `Glass` dataset, `QGE` obtains a $\tau_{q,\texttt{QGE}}$ value of 0.74 ($\pm$0.12), equivalent to using more than 6 random samples. Additional results (reported in the Appendix) confirm that a similar advantage is also found when measuring Spearman's rank correlation (ρ) instead of Kendall's τ.

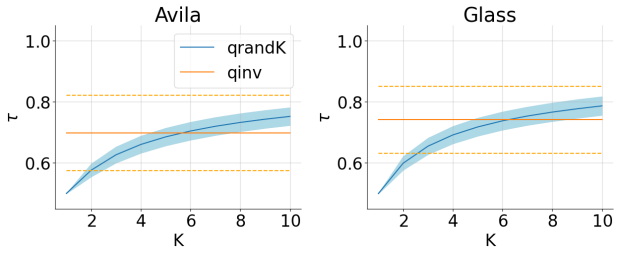


Figure 2: Kendall's τ for the `Avila` and `Glass` datasets. The blue line indicates the average $\tau_{q,\texttt{QRAND}_\texttt{K}}$ for each value of K over 5 different inputs, with the shaded area showing the average $\pm\sigma$. The orange line records the average $\tau_{q,\texttt{QGE}}$, with dashed lines representing the average $\pm\sigma$.)

Ability to rank exceptionally high-quality explanations We analyzed Kendall's τ between q and qt for different subsets of explanations, stratified by their quality levels. The results, depicted in Fig. 3, illustrate that the capability of qt to accurately rank explanations improves with the increasing quality of the explanations. Notably, this improvement is significantly more pronounced for $qt = \texttt{QGE}$ compared to $qt = \texttt{QRAND}_\texttt{K}$, indicating a superior performance in distinguishing high-quality explanations, which are usually the focus of practitioners.

These experiments show that `QGE` meets requirement **R2** to a higher extent than the existing alternative, $\texttt{QRAND}_\texttt{K}$ except for large values of K. Moreover, the computational cost of $\texttt{QRAND}_\texttt{K}$ is proportional to K, so it needs to far exceed the computational cost of `QGE`, which is on par with the computational cost of $\texttt{QRAND}_1$ (two calculations of Ψ). This demonstrates that `QGE` also complies with **R3** to a much higher extent than $\texttt{QRAND}_\texttt{K}$ to obtain comparable performance.

b. Performance on larger datasets To confirm the applicability of our findings across different types of data and larger datasets, we expanded our experiments to include both image datasets (`MNIST`, `CIFAR`, `ImageNet`) and a textual dataset (`20newsgroups`).

For `20newsgroups`, we trained an `MLP` model that achieved an accuracy of 0.78. For `MNIST`, we utilized a small convolutional network with 0.93 accuracy, and for `CIFAR`, a `ResNet50` network obtained an accuracy of 0.77. With `ImageNet`, we tested five different pre-trained

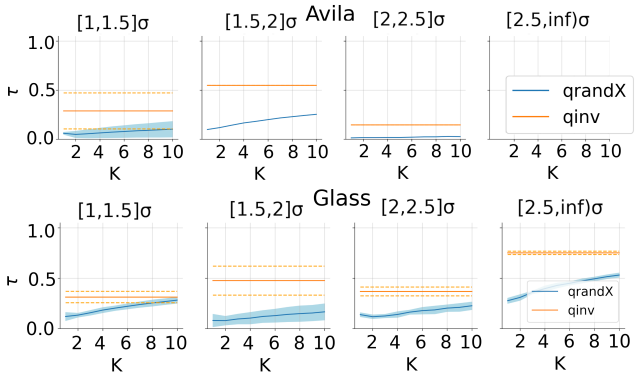


Figure 3: Kendall's τ for explanations of a given level of exceptionality. The blue line represents the average correlation $\tau_{q,\texttt{QRAND}_\texttt{K}}$ for each K value across 10 different inputs, with the shaded area indicating the average$\pm\sigma$. The orange line shows the average correlation $\tau_{q,\texttt{QGE}}$, with dashed lines marking the average $\pm\sigma$.

models from the TorchVision library: three convolutional architectures (`ResNet18`, `ResNet50`, and `VGG16`) and two attention-based architectures (`ViT_b_32` and `MaxViT_t`).

Due to the impracticality of processing all possible explanations for these larger datasets, we sampled 10,000 random explanations for each tested input[2]. For each dataset-model pair, 25 different inputs were tested, and we report the average increase in Kendall's τ when using `QGE` as opposed to $\texttt{QRAND}_1$ ($\Delta\tau = \tau_{q,\texttt{QGE}} - \tau_{q,\texttt{QRAND}_1}$) and the average increase in Spearman's ρ ($\Delta\rho = \rho_{q,\texttt{QGE}} - \rho_{q,\texttt{QRAND}_1}$).

The results[3], detailed in Table 1, consistently show a substantial positive impact on both Kendall and Spearman correlation when using `QGE` instead of $\texttt{QRAND}_1$ across a variety of datasets and model architectures.

c. Advantage of using `QGE` vs. $\texttt{QRAND}_1$ across datasets and models As discussed in Section 4.1, using the `Pixel-Flipping` average activation as a quality measure necessitates a perturbation function that transforms inputs, potentially pushing the model to operate on data points outside of its training distribution (Hase, Xie, and Bansal 2021). To mitigate any effects from this interaction, we re-

[2]Although random explanations frequently have lower quality than explanations obtained with existing explanation methods, the latter are costly to obtain, and too few to yield statistically significant results. Moreover, Section 4.1 shows that the effect observed for average-quality explanations is maintained or enhanced for high-quality explanations, indicating that exploring a sizable set of random explanations is more informative for this experiment than exploring a handful of very good explanations.

[3]Importantly, a well-known issue when explaining predictions involving images is that pixel-level perturbations interact suboptimally with convolutional networks. A prevalent solution involves using explanations at the superpixel level (Blücher, Vielhaben, and Strodthoff 2024)). Therefore, we used superpixel level explanations for `CIFAR` and `ImageNet` (4x4 and 32x32 superpixels, respectively). For completeness, the results for explanations at the pixel level are reported in the Appendix.

DATASET	MODEL	$\Delta\tau$	$\Delta\rho$
20NEWSGROUPS	MLP	0.090±0.05	0.093±0.06
MNIST	MLP	0.214±0.04	0.200±0.03
CIFAR	RESNET50	0.054±0.04	0.059±0.05
IMAGENET	RESNET18	0.110±0.02	0.115±0.02
IMAGENET	RESNET50	0.150±0.03	0.149±0.03
IMAGENET	VGG16	0.108±0.08	0.106±0.07
IMAGENET	VIT_B_32	0.139±0.03	0.141±0.02
IMAGENET	MAXVIT_T	0.132±0.08	0.127±0.07

Table 1: Magnitude of the increase in Kendall and Spearman correlation when using QGE instead of QRAND_1 on large datasets.

AVILA				
MODEL	ACC.	$\sigma(q)$	$\Delta\tau$	$\Delta\rho$
MLP	0.99	0.191	0.198±0.12	0.173±0.10
OOD-MEAN	0.80	0.080	0.288±0.05	0.245±0.03
OOD-ZEROS	0.80	0.085	0.247±0.09	0.213±0.06
UNDERTR.	0.75	0.105	0.315±0.06	0.257±0.03
UNTRAINED	0.05	0.001	0.434±0.01	0.298±0.00
GLASS				
MODEL	ACC.	$\sigma(q)$	$\Delta\tau$	$\Delta\rho$
MLP	0.77	0.198	0.241±0.11	0.204±0.09
OOD-MEAN	0.63	0.070	0.314±0.01	0.262±0.01
OOD-ZEROS	0.63	0.052	0.267±0.06	0.230±0.04
UNDERTR.	0.60	0.168	0.191±0.02	0.184±0.02
UNTRAINED	0.23	0.007	0.173±0.10	0.163±0.08

Table 2: Magnitude of the increase in Kendall and Spearman correlation when using QGE instead of QRAND_1 for the Avila and Glass datasets and different models all using an MLP architecture: MLP refers to the fully trained model exposed to no masked input; ood-mean and ood-zeros were exposed during training to inputs masked with zeros and the average value of each attribute, respectively; undertr. was trained only until achieving 70% the accuracy of the fully trained model; and untrained was not exposed to any data. $\sigma(q)$ indicates the standard deviation of the distribution of q across all possible explanations.

peated the experiments from Section 4.1 using models that were exposed to masked inputs during training, as outlined in (Hase, Xie, and Bansal 2021). Furthermore, we experimented with masking using the average value for each input attribute, rather than zeros. We also tested models with reduced accuracies to diversify the conditions. Table 2 summarizes the average results from these experiments for five different inputs.

These results show that using QGE consistently yields better outcomes than QRAND_1, irrespective of the dataset or model type. However, the nature of the model significantly influences the extent of the advantage offered by QGE. A deeper analysis of this effect is included in Appendix A.4.

PIXEL-FLIPPING		
MODEL	$\Delta\tau$	$\Delta\rho$
RESNET18	0.142±0.07	0.138±0.06
VGG16	0.102±0.04	0.106±0.03
FAITHFULNESSCORRELATION		
MODEL	ΔPA	$\Delta\rho$
RESNET18	0.007±0.01	0.007±0.02
VGG16	0.009±0.01	0.012±0.01
FAITHFULNESSESTIMATE		
MODEL	ΔPA	$\Delta\rho$
RESNET18	0.368±0.02	0.279±0.02
VGG16	0.397±0.02	0.290±0.02
MONOTONICITYCORRELATION		
MODEL	ΔPA	$\Delta\rho$
RESNET18	0.353±0.05	0.277±0.02
VGG1	0.374±0.03	0.288±0.01

Table 3: Magnitude of the increase in Kendall and Spearman correlation when using QGE instead of QRAND_1 for faithfulness metrics on predictions of a ResNet18 and VGG16 models on the Imagenet dataset.

d. Suitability for other quality metrics All experiments reported above use Pixel-Flipping (Bach et al. 2015) as a quality metric. This metric is a popular choice, which is why we have explored it extensively. However, to determine whether the observed effects are consistent across different quality measures, we tested our method using a variety of measures spanning different quality dimensions. For all measures, we used the implementations in the Quantus (Hedström et al. 2023) library.

Faithfulness metrics In addition to Pixel-Flipping, we tested Faithfulness Correlation (Bhatt, Weller, and Moura 2020), Faithfulness Estimate (Alvarez-Melis and Jaakkola 2018), and Monotonicity Correlation (Arya et al. 2019) for 1,000 random explanations. The results, summarized in Table 3, show that QGE performs superiorly to QRAND_1 for all measures, obtaining substantial increases for three of the four measures. However, it offers little advantage for Faithfulness Correlation. This metric is known to be unstable, often yielding highly variable results for the same input across different executions (Tomsett et al. 2020; Hedström et al. 2023; Hedström et al. 2023). This variability undermines the informative advantage of $\Psi(\mathbf{e^{inv}}, \mathbf{x}, f, \hat{y})$ (Eq. 5) over $\Psi(\mathbf{e^r}, \mathbf{x}, f, \hat{y})$ using a random explanation $\mathbf{e^r}$, explaining the lack of advantage observed.

Localization metrics We evaluated the effectiveness of QGE using various localization measures, including AttributionLocalisation (Kohlbrenner et al. 2020), TopKIntersection (Theiner, Müller-Budack, and Ewerth 2022), RelevanceRankAccuracy, RelevanceMassAccuracy (Arras, Osman, and Samek 2022), and AUC (Fawcett 2006). For these tests, we utilized the CMNIST dataset (Bykov et al. 2022), training a

CMNIST - ResNet18		
Measure	$\Delta\tau$	$\Delta\rho$
Attr.Loc.	0.500±0.00	0.309±0.00
TopKInt.	0.015±0.01	0.022±0.01
Rel.RankAcc.	0.089±0.02	0.093±0.02
Rel.MassAcc.	0.501±0.01	0.310±0.01
AUC	0.496±0.00	0.304±0.00

Table 4: Increase in Kendall and Spearman correlation when using QGE instead of QRAND_1 for different localization metrics on predictions of a ResNet18 model on CMNIST data.

ResNet18 model to perform the evaluations. We then assessed the localization of 10,000 random attributions using these measures, applying both the QRAND_1 and QGE transformations. The results are summarized in Table 4, which reports increases in Kendall and Spearman correlation when using QGE instead of QRAND_1. QGE consistently outperforms QRAND_1 across all tested localization metrics, enhancing both the Kendall and the Spearman correlation of the transformed metrics with the original ones. The magnitude of the advantage that QGE provides over QRAND_1 varies depending on the nature of the quality metric used, with the most significant improvements observed using AttributionLocalisation, RelevanceMassAccuracy and AUC.

Robustness and randomization metrics We considered incorporating robustness and randomization metrics into our evaluations. These metrics assess the explanation method itself rather than the explanations obtained. While they can be quantified using the QGE transformation, they are not amenable to comparison using QRAND_1, as the latter does not rely on the explanation method. Therefore, no direct comparison between QGE and QRAND_1 was feasible.

4.2 Effect on Existing Quality Metrics

We investigate the statistical reliability of the QGE transformation compared to the original quality metric. We follow the meta-evaluation methodology outlined in (Hedström et al. 2023) where metric reliability is assessed in two steps: performing a minor- (noise resilience, NR) or disruptive- (reactivity to adversaries, AR) perturbation and then, measuring how the metric scores and method rankings changed, post-perturbation. For each perturbation scenario, intra-consistency (IAC; the similarity of score distributions under perturbation), and inter-consistency (IEC; the ranking similarity among different explanation methods) are computed, resulting in a meta-consistency vector $\mathbf{m} \in \mathbb{R}^4$ and a summarising score $\mathbf{MC} \in [0, 1]$:

$$\mathbf{MC} = \left(\frac{1}{|\mathbf{m}^*|}\right) \mathbf{m}^{*T}\mathbf{m} \quad \text{where} \quad \mathbf{m} = \begin{bmatrix} IAC_{NR} \\ IAC_{AR} \\ IEC_{NR} \\ IEC_{AR} \end{bmatrix}, \tag{8}$$

where $\mathbf{m}^* = \mathbb{1}^4$ represents an optimally performing quality estimator. A higher $\mathbf{MC}$ score, approaching 1, signifies greater reliability on the tested criteria. Perturbations are applied to either the model parameters or input.

fMNIST - LeNet			
Measure	PF	RRA	RMA
QRAND_1	0.579	0.599	0.585
QGE	0.801	0.598	0.587
ImageNet - ResNet18			
Measure	PF	RRA	RMA
QRAND_1	0.634	0.594	0.596
QGE	0.849	0.568	0.619

Table 5: MC score when using QGE and QRAND_1 for Pixel-Flipping (PF) (Faithfulness), Relevance Rank Accuracy (RRA) and Relevance Mass Accuracy (RMA) (Localization) on predictions of a ResNet18 model on the fMNIST dataset.

We used the fMNIST (Xiao, Rasul, and Vollgraf 2017) and ImageNet (Russakovsky et al. 2015) datasets. For the first toy dataset we use the LeNet architecture (LeCun, Cortes, and Burges 2010) and for ImageNet we use a pretrained ResNet-18 (He et al. 2016) from (Paszke et al. 2019). The results shown in Table 5 demonstrate that the QGE method yields reliability improvements across tested metrics. This performance enhancement is most notable for Pixel-Flipping, where QGE significantly enhances the inter-consistency (IEC) under adversarial reactivity (*AR*), indicating a marked improvement in the metric's ability to differentiate between meaningful and random inputs and models, as detailed in Appendix A.5. Also, QGE's effect on localization is competitive, though not statistically significant.

5 Conclusions and Future Work

In this work, we introduced the Quality Gap Estimator (QGE), designed to compare the quality of an explanation against alternative explanations, aiding practitioners in determining the need to seek better alternatives. QGE is computationally efficient and can be used with most quality metrics commonly used in XAI, improving their informativeness.

By conceptualizing the challenge of achieving a relative quality measurement as a sampling issue, we demonstrated that QGE is more sample-efficient than the conventional method of comparing with a single random explanation. Extensive testing across various datasets, models, and quality metrics has consistently shown that employing QGE is advantageous over the traditional approach.

Additionally, the transformation implemented by QGE results in quality metrics with enhanced statistical significance, suggesting its utility even in scenarios where relative comparisons are not the primary objective.

For future work, we aim to enhance QGE's performance with metrics that are inherently unstable, where it currently does not offer a significant improvement over the comparison with a single random sample. Further, we are interested in exploring the potential of employing a similar strategy to also improve the explanations, extending the utility of QGE beyond mere quality measurement.

A Additional Results

This Appendix lists results that complement those mentioned in Section 4.

A.1 Distribution of QGE

The exhaustive exploration of all possible explanations for the 5 input samples used for each of the Avila and Glass datasets confirms that the distribution of QGE is centered around zero, as desired for **R1** and shown in Fig. 4.

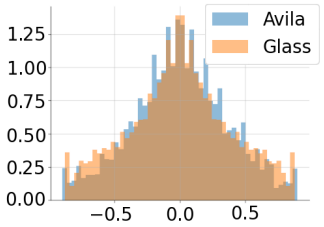


Figure 4: Density histogram of QGE for every possible explanation of 5 different input samples for datasets Avila and Glass.

A.2 Spearman Correlation Results

To complete the analysis in Section 4.1.a we also measured the Spearman correlation ($\rho_{q,qt}$) across the same 5 input samples. Fig. 5 shows that QGE outperforms QRAND_K with up to $k = 4$ samples for Avila, and $k = 5$ for Glass.

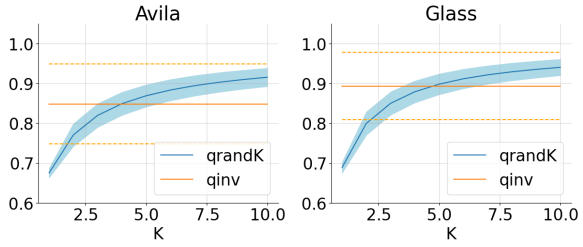


Figure 5: Spearman correlation of qt with the original q for the Avila and Glass datasets. The blue line indicates the average correlation $\rho_{q,\text{QRAND}_\text{K}}$ for each value of K over 5 different inputs. The shaded area shows the average $\pm\sigma$. The orange line records the average correlation $\rho_{q,\text{QGE}}$ with dashed lines representing the average $\pm\sigma$.

A.3 Pixel-Level Explanations

Section 4.1.b reports experiments performed on superpixel-level explanations. For those explanations, instead of perturbing input variables separately, the perturbations are applied to blocks of contiguous pixels, denominated superpixels. The superpixel size used for CIFAR was 4x4, while for ImageNet we used 32x32. For completeness, Table 6 reports the results for perturbations applied to individual pixels, which is analogous to the perturbation mode used in all other datasets. These results show that despite having less of an impact than for super-pixel-level explanations, using QGE is advantageous over QRAND_1. The enhancement of the effect for superpixel-level explanations (which are higher-quality explanations), confirms the result listed in Section 4.1 that indicates that the advantage of QGE is larger the higher the quality of the explanations.

DATASET	MODEL	$\Delta\tau$	$\Delta\rho$
CIFAR	RESNET50	0.054±0.03	0.058±0.03
IMAGENET	RESNET18	0.018±0.01	0.020±0.01
IMAGENET	RESNET50	0.019±0.01	0.021±0.01
IMAGENET	VGG16	0.016±0.01	0.018±0.01
IMAGENET	VIT_B_32	0.040±0.02	0.043±0.02
IMAGENET	MAXVIT_T	0.017±0.02	0.018±0.02

Table 6: Magnitude of the increase in Kendall and Spearman correlation when using QGE instead of QRAND_1 for explanations at the pixel level.

A.4 Variation of the Effect With the Distribution's Standard Deviation

A deeper analysis of the results in Section 4.1.c shows considerable variation in the distribution of q depending on the model used, as illustrated in Figure 6. Table 2 also presents the average standard deviation of the q distribution for each model. The fully-trained *mlp* model exhibits a wide range of q values for each input (the average σ is 0.19), indicating a substantial difference in quality between the best and worst explanations. In contrast, the *undertrained* model displays significantly less variability in q across both datasets. The most pronounced case is the *untrained* model, which shows highly concentrated q distributions, i.e., minimal numerical differences between the q values of the best and worst explanations. Despite these variations, the impact of utilizing QGE over QRAND_1 is consistently positive, confirming its robustness and effectiveness across various conditions.

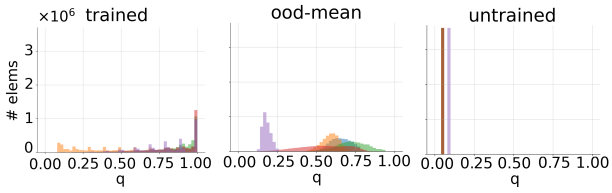
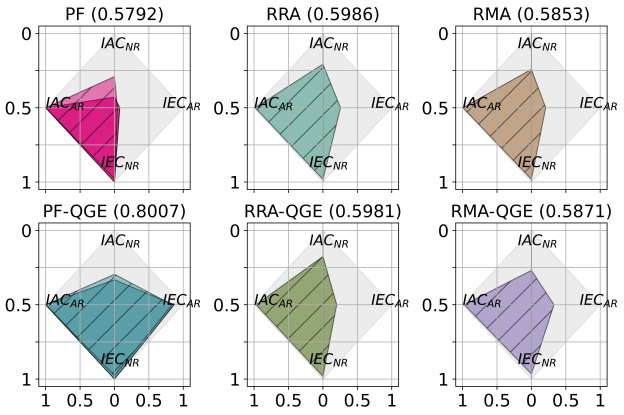


Figure 6: Histograms for the distributions of q for 5 different inputs using the *trained*, *ood-mean* and *untrained* models.

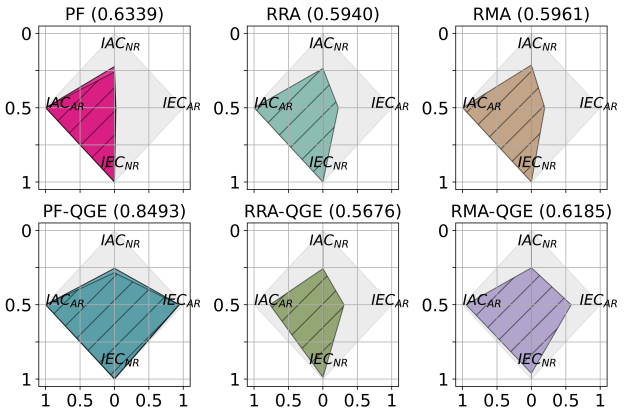
A.5 Effect on Existing Quality Measures

In Figure 7, we show the different area graphs which each contain the results from the meta-evaluation analysis (set as coordinates on a 2D plane) for the fMNIST (Xiao, Rasul, and Vollgraf 2017) and ImageNet (Russakovsky et al. 2015) datasets, respectively. The titles hold the summarising **MC** score and each edge contains the meta-evaluation vector scores. By inspecting the colored areas of the respective

estimators in terms of their size and shape, we can deduce the overall performance of both failure modes. Larger colored areas imply better performance on the different scoring criteria and the grey area indicates the area of an optimally performing quality estimator. The Quantus (Hedström et al. 2023) and MetaQuantus (Hedström et al. 2023) libraries were used for the experiments.



(a) `fMNIST - LeNet`



(b) `ImageNet - ResNet18`

Figure 7: A graphical representation of the benchmarking results aggregated over 3 iterations with $K = 5$. We use {*Saliency, Integrated Gradients, Input X Gradient*} as explanation methods. Each column corresponds to a quality estimator, from left to right: `Pixel-Flipping` (PF) (Faithfulness), `Relevance Rank Accuracy` (RRA) and `Relevance Mass Accuracy` (RMA) (Localization). The bottom row shows results with `QGE`. Solid shapes correspond to input perturbations, and striped shapes to model perturbations. 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 Model- and Input perturbation tests. Higher values and larger colored areas indicate higher performance.

B Code and Reproducibility

An implementation of `QGE` for a wide variety of quality measures is available in the Quantus toolkit: https://github.com/understandable-machine-intelligence-lab/Quantus

The code used for all experiments is available at https://github.com/annahedstroem/eval-project

Acknowledgments

Carlos Eiras-Franco wishes to thank CITIC, which financed the research stay that originated this work. As a Research Center accredited by the Galician University System, CITIC is funded by "Consellería de Cultura, Educación e Universidades from Xunta de Galicia", supported in an 80% through ERDF Funds, ERDF Operational Programme Galicia 2014-2020, and the remaining 20% by "Secretaría Xeral de Universidades" (Grant ED431G 2019/01). This publication is part of project PID2021-128045OA-I00, financed by MCIN/AEI/10.13039/501100011033/FEDER, UE. Carlos Eiras-Franco also thanks the support received by the Xunta de Galicia (Grant ED431C 2018/34) with the European Union ERDF funds, and the support received by the National Plan for Scientific and Technical Research and Innovation of the Spanish Government (Grant PID2019-109238GBC22) and the European Union ERDF funds. This work was funded by the German Ministry for Education and Research through project Explaining 4.0 (ref. 01IS200551).

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I. J.; Hardt, M.; and Kim, B. 2018. Sanity Checks for Saliency Maps. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 9525–9536.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 7786–7795.

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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

Arya, V.; Bellamy, R. K. E.; Chen, P.; Dhurandhar, A.; Hind, M.; Hoffman, S. C.; Houde, S.; Liao, Q. V.; Luss, R.; Mojsilovic, A.; Mourad, S.; Pedemonte, P.; Raghavendra, R.; Richards, J. T.; Sattigeri, P.; Shanmugam, K.; Singh, M.; Varshney, K. R.; Wei, D.; and Zhang, Y. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear

classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7).

Bhatt, U.; Weller, A.; and Moura, J. M. F. 2020. Evaluating and Aggregating Feature-based Model Explanations. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3016–3022. ijcai.org.

Blücher, S.; Vielhaben, J.; and Strodthoff, N. 2024. Decoupling Pixel Flipping and Occlusion Strategy for Consistent XAI Benchmarks. arXiv:2401.06654.

Bommer, P.; Kretschmer, M.; Hedström, A.; Bareeva, D.; and Höhne, M. M. 2023. Finding the right XAI method - A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science. *CoRR*, abs/2303.00652.

Brocki, L.; and Chung, N. C. 2022. Evaluation of Interpretability Methods and Perturbation Artifacts in Deep Neural Networks. *CoRR*, abs/2203.02928.

Brunke, L.; Agrawal, P.; and George, N. 2020. Evaluating Input Perturbation Methods for Interpreting CNNs and Saliency Map Comparison. In *Computer Vision – ECCV 2020 Workshops*, 120–134. Springer International Publishing.

Bykov, K.; Hedström, A.; Nakajima, S.; and Höhne, M. M. 2022. NoiseGrad - Enhancing Explanations by Introducing Stochasticity to Model Weights. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 6132–6140. AAAI Press.

Dasgupta, S.; Frost, N.; and Moshkovitz, M. 2022. Framework for evaluating faithfulness of local explanations. In *International Conference on Machine Learning*, 4794–4815. PMLR.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8): 861–874.

German, B. 1987. Glass Identification. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5WW2P.

Hase, P.; Xie, H.; and Bansal, M. 2021. The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 3650–3666.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. 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*, 770–778. IEEE Computer Society.

Hedström, A.; Weber, L.; Krakowczyk, D.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; and Höhne, M. M.-C. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, 24(34): 1–11.

Hedström, A.; Bommer, P.; Wickstrøm, K. K.; Samek, W.; Lapuschkin, S.; and Höhne, M. M. C. 2023. The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus. *arXiv preprint arXiv:2302.07265*.

Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; and Lapuschkin, S. 2020. Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Nguyen, A.; and Martinez, M. R. 2020. On quantitative aspects of model interpretability. *CoRR*, abs/2007.07584.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Rieger, L.; and Hansen, L. K. 2020. IROF: a low resource evaluation metric for explanation methods. *CoRR*, abs/2003.08747.

Rong, Y.; Leemann, T.; Borisov, V.; Kasneci, G.; and Kasneci, E. 2022. A Consistent and Efficient Evaluation Strategy for Attribution Methods. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 18770–18795. PMLR.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.*, 115(3): 211–252.

Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; and Müller, K. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans. Neural Networks Learn. Syst.*, 28(11): 2660–2673.

Stefano, C.; Fontanella, F.; Maniaci, M.; and Freca, A. 2018. Avila. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K02X.

Theiner, J.; Müller-Budack, E.; and Ewerth, R. 2022. Interpretable Semantic Photo Geolocation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, 1474–1484. IEEE.

Tomsett, R.; Harborne, D.; Chakraborty, S.; Gurram, P.; and Preece, A. D. 2020. 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*, 6021–6029. AAAI Press.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, abs/1708.07747.