# Evaluating Model Explanations without Ground Truth

Kaivalya Rawal
kaivalyarawal45@gmail.com
Oxford Internet Institute
Oxford, Oxfordshire, UK

Zihao Fu
zihao.fu@oii.ox.ac.uk
Oxford Internet Institute
Oxford, Oxfordshire, UK

Eoin Delaney
eoin.delaney@tcd.ie
Trinity College Dublin
Dublin, Ireland

Chris Russell
chris.russell@oii.ox.ac.uk
Oxford Internet Institute
Oxford, Oxfordshire, UK



## Abstract

There can be many competing and contradictory explanations for a single model prediction, making it difficult to select which one to use. Current explanation evaluation frameworks measure quality by comparing against ideal "ground-truth" explanations, or by verifying model sensitivity to important inputs. We outline the limitations of these approaches, and propose three desirable principles to ground the future development of explanation evaluation strategies for local feature importance explanations. We propose a ground-truth **Agnostic eXplanation Evaluation** framework (**AXE**) for evaluating and comparing model explanations that satisfies these principles. Unlike prior approaches, AXE does not require access to ideal ground-truth explanations for comparison, or rely on model sensitivity – providing an independent measure of explanation quality. We verify AXE by comparing with baselines, and show how it can be used to detect explanation fairwashing. Our code is available at https://github.com/KaiRawal/Evaluating-Model-Explanations-without-Ground-Truth.

## CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

## Keywords

explainability, interpretability, XAI, evaluation, benchmark

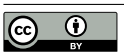
**ACM Reference Format:**
Kaivalya Rawal, Zihao Fu, Eoin Delaney, and Chris Russell. 2025. Evaluating Model Explanations without Ground Truth. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3715275.3732219

## 1 Introduction

As artificial intelligence (AI) systems are increasingly used in critical decision-making processes, knowing which model explanation to trust has emerged as a fundamental challenge. Model explanations often disagree with each other (see figure 1), and the selection of incorrect or intentionally misleading explanations can have far-reaching consequences – from misinforming users and regulators to reinforcing systemic biases and eroding public trust in AI systems [5, 37, 39]. This challenge is particularly acute in high-stakes domains like healthcare diagnostics, financial and credit scoring services, and criminal justice, where machine learning models directly impact human lives. It is essential to be able to select the best explanation from a set of possible explanations, but unfortunately there has been little progress towards this critical problem.

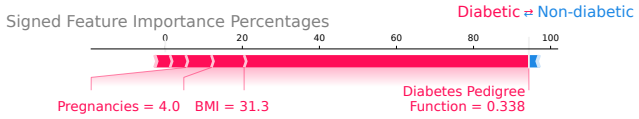
User studies offer a workaround – of approximately 300 papers proposing new model explanation methods (explainers), one in seven performed user study evaluations [42]. However, one in three papers evaluated entirely anecdotally, reflecting the need for standardized evaluation frameworks [42]. Without consensus on the essential properties that explanations should possess and robust frameworks to numerically evaluate them, progress in the field remains slow and fragmented. Historically, advances in AI have often been driven by benchmark datasets and deterministic evaluation frameworks defined using standard metrics – as exemplified by ImageNet for computer vision [15] and MMLU for language understanding [26]. Developing analogous benchmarks for eXplainable AI (XAI) involves unique challenges, including a lack of access to reliable "ground-truth" explanations to compare against. This hampers our ability to meaningfully evaluate competing explanation methods, assess their utility to users impacted by AI systems, or their faithfulness to model behavior. Progress towards these goals can ensure that XAI truly makes AI systems transparent.

There are many forms of model explanation. One popular category among practitioners is post-hoc model-agnostic feature-importance explanations, such as LIME [48] or SHAP [36]. These provide explanations for individual predictions rather than describing global model behavior. They can operate on any model type, including neural networks, regardless of weights or architecture. They produce feature importances as output: a signed vector indicating the relative contribution of each input feature to the output. While there are many competing explanation types, data modalities, and evaluation desiderata, [33], this paper focuses exclusively on local feature-importance explanations for models operating on tabular datasets. Even in this restricted setting, different explanation methods (explainers) often provide contradictory explanations (figure 1). In this paper we do not propose a new XAI method but instead develop three general principles: *local contextualization*, *model relativism*, and *on-manifold evaluation* to guide the evaluation of feature-importance explanations. We use these to propose AXE, a

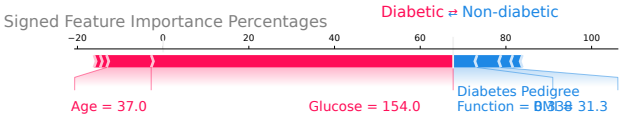



*FAccT '25, Athens, Greece*

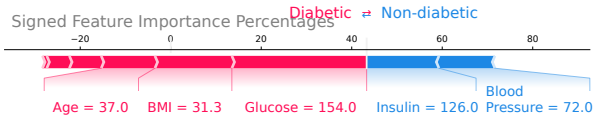
ACM ISBN 979-8-4007-1482-5/2025/06
https://doi.org/10.1145/3715275.3732219



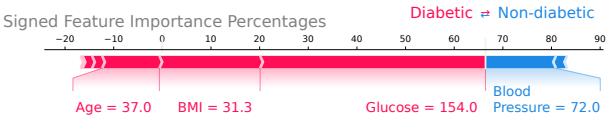

**(a) "Gradients" explainer: Diabetes Pedigree Function is the most important input feature, pushing the classifier to a positive prediction (diabetic)**




**(b) "SHAP" explainer: Glucose is still the most important (positive) input, BMI now has an negative importance (non-diabetic)**




**(c) "LIME" explainer: Glucose (positive), BMI (positive), and Insulin (negative) are all important input features**




**(d) "Integrated Gradients" explainer: Glucose (positive) BMI (positive) and Blood Pressure (negative) are important.**

**Figure 1: Different Explainers Yield Different Explanations: A neural network predicts diabetes on the "Pima Indians" dataset [10]. A single positive (diabetic) prediction is explained using four explainers. These feature-importance explanations, visualized here as "force-plots", consist of a signed vector indicating the relative contribution of each input to the model output. They disagree with each other. Section 2.3 details the explainers, and section 3.1 evaluates these four explanations using AXE.**

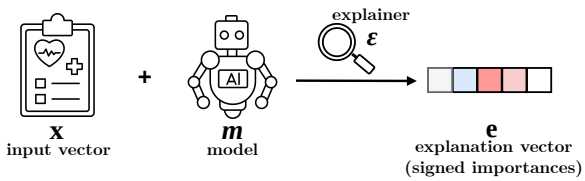
new ground-truth **Agnostic eXplanation Evaluation** framework that considers a good explanation to be one that correctly identifies the features most predictive of model outputs. AXE is inspired by user research which indicates useful explanations are those that help users emulate and predict model behavior [12].

The plots in figure 1 visualize competing explanations for the same datapoint in a diabetes classification model. They disagree with each other in the contributions of the input features, a phenomenon commonly documented in the literature [5, 31, 39]. Some XAI methods such as LIME and SHAP often rely on off-manifold model predictions to generate a single explanation, leading to different explanations. This explicit reliance on feature sensitivity is one potential cause for explanation disagreement, which we seek to address through the *on-manifold evaluation* principle proposed in section 2.2 A good explanation evaluation framework should provide clear guidance about which explanation is better, helping users make sense of competing explainers. Explanation disagreement can be exploited by adversaries to produce fairwashed explanations – where a given explainer certifies that protected attributes were not important to the model even if they determined the prediction [3, 54]. This presents a risk for auditors and regulators enforcing AI fairness, further motivating our work and highlighting the importance of evaluating explanation quality.

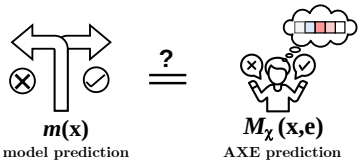
This paper is structured as follows: in section 2 we define our notation, introduce three foundational principles for explanation evaluation, and describe prior work. In section 3 we introduce AXE, an explanation evaluation framework directly couched in terms of predictive accuracy – the notion that a good human-interpretable explanation is one which identifies the features most predictive of the model behavior [11, 12]. In section 4, we demonstrate how AXE can be used to detect explanation fairwashing – foiling a state-of-the-art adversarial attack [54], and compare AXE with existing baselines from the literature. We conclude in section 5 with a brief summary discussion.

## 2 Evaluating Model Explanations

A typical scenario depicting the generation and evaluation of local, post-hoc, model-agnostic explanations is presented in figure 2.




**(a) Explanation generation: Explainer $\mathcal{E}$ produces an explanation vector e of signed feature importances using datapoint x, model $m$ and prediction $m(\mathbf{x})$**




**(b) Explanation evaluation: AXE evaluates the quality $q$ of explanation e by measuring how accurately prediction $m(\mathbf{x})$ can be recovered from dataset $\mathcal{X}$.**

**Figure 2: Explanation Generation (a) and Evaluation (b): AXE measures how well a given explanation can help emulate model behavior. See section 3.1 for full algorithm.**

### 2.1 Our Notation

We specify our notation from figure 2: input vector $\mathbf{x}$, model $m$, and explanation $\mathbf{e}$. We use these to define an explanation "quality" metric $q$ and an evaluation framework $Q$ here, and in algorithm 1 we implement such a framework using AXE.

(1) *Input Vector*: The input feature vector for any arbitrary datapoint is defined as: $\mathbf{x} = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^N$, where $N$ is the number of features, and each $x_i$ is a real-valued feature.

(2) *Model and Prediction*: The model $m$ is a mapping from the feature space to a binary output: $m : \mathbb{R}^N \to \{0, 1\}$, and the prediction for input $\mathbf{x}$ is given by: $m(\mathbf{x}) = y_{\text{pred}} \in \{0, 1\}$.
(3) *Explanation*: A local feature importance explanation is denoted $\mathbf{e}$. It is a function of the input $\mathbf{x}$ and model $m$ (implicitly model prediction $m(\mathbf{x})$ too). For an explainer $\mathcal{E}$, $\mathbf{e} = \mathcal{E}(\mathbf{x}, m)$, where $\mathbf{e} \in \mathbb{R}^N$, and each component $e_i$ represents the (signed) contribution or importance of the feature $x_i$ to the prediction $m(\mathbf{x})$.
(4) *Explanation Quality Metric*: For dataset $\mathcal{X}$, the explanation quality metric $q \in [0, 1]$ evaluates the quality of explanation $\mathbf{e}$ for a specific input $\mathbf{x}$ and model $m$. As a function, $q = q_{\mathcal{X}}(\mathbf{x}, m, \mathbf{e})$ where $0 \leq q \leq 1$ (greater $q$ is better). Previous work often refers to quality scores as fidelity or explanation faithfulness [8, 23, 35].

An *explanation evaluation framework* is a tuple $(\mathcal{X}, m, \mathcal{E}, Q)$:

- $\mathcal{X} \in \mathbb{R}^{v \times N}$ is the dataset of inputs with $N$ features and $v$ datapoints, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_v\}$.
- $m : \mathbb{R}^N \to \{0, 1\}$ is the model being explained.
- $\mathcal{E} : \mathbb{R}^N \to \mathbb{R}^N$ is the explanation method that generates explanation $\mathbf{e} \in \mathbb{R}^N$ for each datapoint $\mathbf{x} \in \mathbb{R}^N$.
- $Q$ is the aggregate quality score over the dataset computed as an average of explanation quality $q$:

$$Q(\mathcal{X}, m, \mathcal{E}) = \frac{1}{v} \sum_{i=1}^{v} q(\mathbf{x}_i, m, \mathcal{E}(\mathbf{x}_i, m), \mathcal{X}).$$

## 2.2 Three Principles for Evaluating Model Explanation Quality

Variations in explanations occur for many reasons. For example: (a) different input datapoints $\mathbf{x}_1 \neq \mathbf{x}_2$ typically have different explanations; (b) different prediction models $m_1 \neq m_2$ – eg. with updated neural network weights – typically have different explanations; and (c) as seen in figure 1, explanations from different explainers can have different explanations, possibly due to varying off-manifold input sensitivity of the model $m$ in the neighborhood of $\mathbf{x}$. An evaluation framework that cannot distinguish between explanations from these varying scenarios and always scores diverse explanations the same is not helpful. We characterize these situations respectively with the following principles:

(1) **Local Contextualization**: *Explanations should depend on the datapoint being explained.* For local explanations, when the datapoint $\mathbf{x}$ changes, the evaluation metric $q$ should not always prefer that the corresponding explanation $\mathbf{e}$ remain unchanged. Model behavior is not always identical across the data distribution.
(2) **Model Relativism**: *Explanations should depend on the model being explained.* When the model $m$ changes, the evaluation metric $q$ should not always prefer that the corresponding explanation $\mathbf{e}$ remain unchanged.
(3) **On-manifold Evaluation**: *Explanations on-manifold should not depend on changes in off-manifold model behavior.* When off-manifold model predictions $m(\mathbf{x} + \delta\mathbf{x})$ change, the evaluation metric $q$ should remain unchanged for explanation $\mathbf{e}$. Evaluation metrics should not make the same assumptions as the explainers they seek to evaluate – which often assume changes in output caused by synthetic perturbations in particular model inputs indicate the importance of those inputs.

The *on-manifold evaluation* principle is motivated by the observation that many explanation methods are variants of sensitivity analysis that capture how much synthetically varying a particular feature alters model outputs [17, 29, 46]. Ideally, an explanation for model behavior on datapoint $\mathbf{x}_1$ should not depend on model

**Table 1: Explanation Evaluation Metrics: Definitions for ground-truth based explanation evaluation metrics: FA, RA, SA, SRA, RC and PRA [2, 31] (e is an explanation, and $\mathbf{e}^*$ is the ground truth); sensitivity based metrics PGI and PGU [2, 14, 43]; and AXE. For each we list whether it satisfies the three evaluation principles laid out in section 2.2. * *For PGU, lower values are better***

| Metric | Definition | Local Contextualization | Model Relativism | On-Manifold Evaluation |
|---|---|---|---|---|
| **FA:** Feature Agreement | Fraction of top-n features common between $\mathbf{e}$ and $\mathbf{e}^*$. | ✗ | ✗ | ✔ |
| **RA:** Rank Agreement | Fraction of top-n features common between $\mathbf{e}$ and $\mathbf{e}^*$ with the same position in respective rank orders. | ✗ | ✗ | ✔ |
| **SA:** Sign Agreement | Fraction of top-n features common between $\mathbf{e}$ and $\mathbf{e}^*$ with the same sign. | ✗ | ✗ | ✔ |
| **SRA:** Signed Rank Agreement | Fraction of top-n features common between $\mathbf{e}$ and $\mathbf{e}^*$ with the same sign and rank. | ✗ | ✗ | ✔ |
| **RC:** Rank Correlation | Spearman's rank correlation coefficient for feature rankings from $\mathbf{e}$ and $\mathbf{e}^*$. | ✗ | ✗ | ✔ |
| **PRA:** Pairwise Rank Agreement | Fraction of feature pairs for which relative ordering in $\mathbf{e}$ and $\mathbf{e}^*$ is the same. | ✗ | ✗ | ✔ |
| **PGI:** Prediction-Gap on Important Feature Perturbation | Mean absolute change in model output upon perturbing top-n most important inputs. | ✔ | ✔ | ✗ |
| **PGU*:** Prediction-Gap on Unimportant Feature Perturbation | Mean absolute change in model output upon perturbing top-n most unimportant inputs. | ✔ | ✔ | ✗ |
| **AXE:** (ground-truth) Agnostic eXplanation Evaluation | Predictiveness of the top-n most important inputs in recovering model output. **Defined in section 3.1.** | ✔ | ✔ | ✔ |

behavior on a different datapoint $\mathbf{x_2} = \mathbf{x} + \delta\mathbf{x}$. Further, evaluation frameworks that capture the fidelity of explanations with respect to synthetic neighborhoods around real points, are simply encoding a particular choice of sensitivity analysis without meaningfully evaluating the explanation quality. Section 2.5 formalizes this.

Previous methods for computing the quality $q$ of explanation $\mathbf{e}$ have suggested comparing $\mathbf{e}$ with a known "ground-truth" vector $\mathbf{e}^*$. Proposals include one "ground-truth" per datapoint $\mathbf{x}$, unintentionally introducing independence from $m$ [1, 57]; or one "ground-truth" per model $m$, introducing independence from $\mathbf{x}$ [2, 31]. The latter case clearly violates *local contextualization* by comparing each local explanation with the same static "ground-truth", promoting a holistic global model explanation instead of local explanations that differ across datapoints. The former case violates *model relativism* by computing quality $q$ for explanation $\mathbf{e}$ using an immutable "ground-truth" $\mathbf{e}^*$, fixed for a given datapoint, regardless of the model used. With images especially, an explanation is often considered good if it selects the "correct" region as important in an image – regardless of whether the model used those features [1, 22, 57]. Section 2.4 showcases these violations in detail.

## 2.3 Prior Approaches

Explanations can be evaluated using any of the evaluation metrics defined in table 1. Broadly, these fall into two categories:

(1) *ground-truth* based metrics compare the generated explanations $\mathbf{e}$ with ground-truth annotations $\mathbf{e}^*$, either collected by humans or inferred using a different proxy [20, 21, 44, 51]. These include Feature Agreement (FA), Sign Agreement (SA), Rank Agreement (RA), Signed Rank Agreement (SRA), Rank Correlation (RC) and Pairwise Rank Agreement (PRA) [31].
(2) *sensitivity* based metrics verify model sensitivity to the inputs declared important by an explanation [17, 30, 44, 50]. These have been summarized as Prediction Gap on Important Feature Perturbation (PGI) and Prediction Gap on Unimportant Feature Perturbation (PGU) [2, 14, 43].

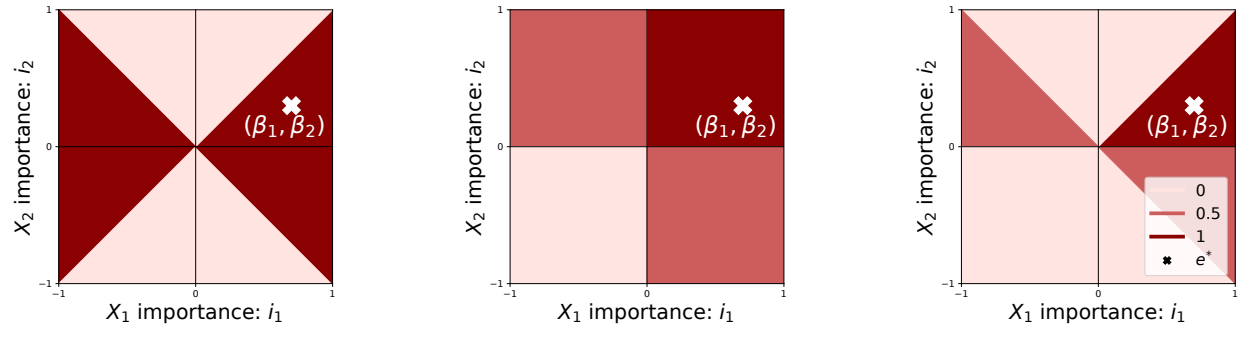
In addition to evaluation metrics $q$, we also summarize the most common explanation methods (explainers) $\mathcal{E}$. We limit ourselves to post-hoc explainers that produce signed feature importance vectors as explanations. Many of these are inspired by sensitivity analysis – measuring how changes in input variables effect changes in the model response [17, 29, 46]. Both LIME and SHAP sample synthetic points in the neighborhood of a given datapoint $\mathbf{x}$ and fit linear models to obtain feature importances [36, 48]. Gradient-based methods compute the gradient of the model output with respect to the input [53], with several extensions: Smooth Grad [55], Integrated Gradients [56], and Input × Grad [52]. In our experiments in section 4.2, we used the standard OpenXAI benchmark for generating explanations $\mathbf{e}$ and evaluating them using prior metrics $q$. [2].

Several user studies have demonstrated that for explanations to be useful in real-world scenarios, their primary function must be to help users predict model behavior [4, 7, 9, 12, 25, 45]. AXE is designed to explicitly operationalize this idea, filling a critical gap in the literature. While work has referenced the need to move beyond ground-truth and sensitivity towards predictiveness as a measure of explanation quality [11, 16, 34], few metrics have implemented this idea. Some previous instantiations have been used to measure the quality of explanations in image classification [22, 28], where *model relativism* violations are common and egregious [1, 22, 57].

## 2.4 Invariance of ground-truth Metrics to Changing Data and Models

Real world situations lack access to an oracle to provide ground-truth explanations $\mathbf{e}^*$ [41, 58]. For linear models, a common resolution adopts the model coefficients as the "ground-truth" for all datapoints $\mathbf{x} \in \mathcal{X}$ in a dataset [2, 31]. As mentioned in section 2.2, comparing with the same $\mathbf{e}^* \forall \mathbf{e} \in E$ is undesirable for local explanations and promotes a single explanation across datapoints.

Figure 3 depicts such an example, directly violating the *local contextualization* principle. Consider a model with two input features, $X_1$ and $X_2$, and prediction $y$, parameterized $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. For datapoint $\mathbf{x}$ feature importances are $i_1$ and $i_2$, with explanation $e = [i_1, i_2]$. FA, RA, SA, SRA, RC, and PRA measure explanation quality by comparing with ground-truth explanation $\mathbf{e}^* = [\beta_1, \beta_2]$ [31]. Since $\mathbf{e}^*$ is constant and independent of $\mathbf{x}$, every explanation is compared against the same tuple $[\beta_1, \beta_2]$. This comparison takes many forms, with definitions provided in table 1. For




(a) Rank Agreement: $RA_{n=2}$ (b) Sign Agreement: $SA_{n=2}$ (c) Signed Rank Agreement: $SRA_{n=2}$

**Figure 3: Violations of local contextualization and model relativism: Plots showing explanation quality $q$ (color) across $i_1$ and $i_2$ values for explanation $\mathbf{e} = (i_1, i_2)$. Model $m(\mathbf{x}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ has ground-truth $\mathbf{e}^* = (\beta_1, \beta_2) = (0.7, 0.3)$. Diverse explanations $\mathbf{e}$ map to the same quality $q$ (0, 0.5, or 1), violating *local contextualization*. Changing the model changes the ground-truth $\mathbf{e}^*$, but leaves the plots unchanged $\forall \beta_1, \beta_2$ where $\beta_1 > \beta_2 > 0$, violating *model relativism*. Section 2.4 explains these computations.**

example, our $N = 2$ feature setup implies that for the top $n$ features: $\text{FA}_{n=0} = 0$, $\text{FA}_{n=1} \in \{0, 0.5, 1.0\}$, and $\text{FA}_{n=2} = 1$, and that $\text{FA}_{n=1} = \text{RA}_{n=2} = \text{PRA}_{n=2}$, while RC is undefined.

Plotting evaluation metric $q$ for all possible explanations $e = [i_1, i_2]$, for an example model with $\beta_1 = 0.7$ and $\beta_2 = 0.3$, we see that regardless of the specific value of $\mathbf{e}$, there are regions where the resulting $RA_{n=2}$ is the same (figure 3 a). Similarly, the $SA_{n=2}$ is the same across $i_1, i_2$ regions (figure 3 b) and $SRA_{n=2}$ too (figure 3 c). Concretely, any explanation $e = (i_1, i_2)$ such that $i_1 > 0$, $i_2 > 0$, and $i_1 > i_2$ (this is the region labeled 1 in figure 3 c) is guaranteed to have the same FA, RA, SA, SRA, RC, and PRA. These plots are specific to our particular model and ground-truth, but display multiple regions of constant FA, RA, SA, SRA, and PRA values, displaying a *violation of the local contextualization principle*. In real world settings, these metrics would fail to distinguish different explanations from each other in quality if they belonged to the same region in figure 3.

The same logic demonstrates violations of *model relativism*. From table 1 we can see that all ground-truth comparison metrics are symmetric. The metrics are invariant to changes in $\mathbf{e}^*$, the same way they are invariant to changes in $\mathbf{e}$. For a given explanation $\mathbf{e}$, and for any model $m$ such that $\beta_1 > \beta_2 > 0$, the plots in figure 3 would stay unchanged. Concretely, while we used model weights (ignoring $\beta_0$) $m = (\beta_1, \beta_2) = (0.7, 0.3)$, the plots would be unchanged for $m_a = (0.99, 0.01)$, $m_b = (0.99, 0.98)$, and $m_c = (0.02, 0.01)$ – vastly different models! The importance of $X_1$ with respect to $X_2$ ranges from 1 to $\infty$ in the limit, and it is absurd for an explanation quality metric $q$ to be unchanged for these diverse models. In this way, FA, RA, SA, SRA, RC, and PRA *fail the model relativism principle*, no longer distinguishing explanations by quality when the underlying model changes.

## 2.5 Inherent biases in sensitivity Metrics

Sensitivity based explainers $\mathcal{E}$ like LIME are highly sensitive to hyperparameters. This facilitates adversarial fairwashing attacks (section 4.1) [54] and can cause feature importances to switch arbitrarily from highly positive to highly negative [40, 49]. Sensitivity analysis based evaluation metrics $q$ suffer similar problems.

Metrics like PGI and PGU may simply encode a preference for particular explainers. We formalize this in the context of synthetic data. Consider a wide range of explainability measures that measure some loss, $\ell$, defined in terms of the fidelity $F$ to classifier responses $m(\cdot)$, in a synthetic neighborhood $\mathcal{N}_\mathbf{x}$ around each datapoint $\mathbf{x}$.

$$\ell = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{n \in \mathcal{N}_\mathbf{x}} F(m(n), \hat{c}_\mathbf{x}(n)) \tag{1}$$

$\hat{c}_\mathbf{x}(n)$ is typically defined as something analogous to a first-order Taylor expansion about $\mathbf{x}$, taking the form $\hat{c}_\mathbf{x}(n) = m(\mathbf{x}) + I \cdot (n - \mathbf{x})$, where instead of $I$ being the gradient of function $m$, it is the per-datapoint and per-feature importance returned by explainer $\mathcal{E}$.

However, as per datapoint feature importance is typically computed by fitting a simple linear model over the synthetic points [48], we can simply consider a new feature-importance explanation method given by the per-point minimizer, thereby matching the explanation evaluation metric exactly:

$$I'(\mathbf{x}) = \arg\min_I \sum_{n \in \mathcal{N}_\mathbf{x}} F(m(n), m(\mathbf{x}) + I \cdot (n - \mathbf{x})) \tag{2}$$

By definition, this is an optimal minimizer of (1), and will perform best with respect to the metric. As trivial examples of this: When $F$ is the squared loss, if $\mathcal{N}$ is defined in terms of homogeneous Gaussian noise then (2) corresponds to the definition of LIME [39]; as the variance of the Gaussian tends to 0, it corresponds to the gradient of the function; and it corresponds to SHAP, if $\mathcal{N}$ is chosen as weighted sampling over the vertices of a cube formed by swapping the values of a particular datapoint $p$ with the distribution mean.

Existing sensitivity based metrics such as PGI define $F$ using $L_1$ loss, with neighborhood $\mathcal{N}$ defined using a Gaussian distribution around datapoint $\mathbf{x}$. While the loss is $L_1$ and not $L_2$, this formulation is otherwise interchangeable with LIME, and also converges to the gradient as the variance of the Gaussian tends to 0. This naturally promotes explainers that satisfy this definition of fidelity and neighborhood, *violating the on-manifold evaluation principle*.

# 3 Methodology

Inspired by previous work (section 2.3) and desiderata from user studies [4, 7, 9, 12, 25, 45], we consider a simple alternative to sensitivity-driven methods of XAI evaluation: the important features in any explanation should be more predictive of the model output than the unimportant features. AXE adopts classifier accuracy [28] to measure the predictiveness of the top-n important features. This "top-n" style formulation, just like prior metrics from table 1, is considered intuitive for practitioners [2].

For datapoint $\mathbf{x}$ and explanation $\mathbf{e}$, the top-n most important features are the importances with the largest absolute values. To measure explanation quality $q$, AXE uses predictiveness – the accuracy of a k-Nearest Neighbors ($k$-NN) model $M^k$ in recovering the model prediction $m(\mathbf{x})$ using only the subset of the the top-n most important features. The $k$-NN $M$ mimics the prediction $m(\mathbf{x})$ of model $m$ by averaging over the predictions from the $k$ neighbors nearest to $\mathbf{x}$ [13, 19]. We motivate our choice of $k$-NN based on feature separability in section 3.2.

## 3.1 The AXE Framework

**Algorithm 1** Evaluating Explanation Quality with $\text{AXE}^{\boldsymbol{k}}_{\boldsymbol{n}}$

**Require:** Number of Features $n$, Number of Neighbors $k$
Dataset $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{v}$, Predictions $Y_{\text{preds}} = \{y_i\}_{i=1}^{v}$, and Explanations $E = \{\mathbf{e}_i\}_{i=1}^{v}$

Initialize an empty list: $\hat{Y} \leftarrow []$
**for** each datapoint $\mathbf{x}_i$ and explanation $\mathbf{e}_i$ in $(\mathcal{X}, E)$ **do**
  Find $n$ most important features: $f_{\text{imp}} \leftarrow \text{ImpFeatures}(\mathbf{e}_i, n)$
  Create $\mathcal{X}_f$ with subset of features $f_{imp}$ from $\mathcal{X}$
  Train K-NN model $M_i^k$ with inputs $\mathcal{X}_f$ and target $Y_{\text{preds}}$
  Obtain prediction $\hat{y}_i$ from $M_i^k$ for datapoint $\mathbf{x}_i$
  Append $\hat{y}_i$ to $\hat{Y}$
**end for**
Return performance measure: $\text{Accuracy}(\hat{Y}, Y_{\text{preds}})$

Algorithm 1 summarizes our framework for evaluating explanations. We denote the target variable $Y$, predicted from the input dataset $\mathcal{X}$ consisting of $v$ datapoints and $N$ features. The model $m$ makes predictions $Y_{\text{preds}} = m(\mathcal{X})$, for which a set of feature-importance explanations $E$ can be computed such that $\exists \mathbf{e}_i \in E \forall \mathbf{x}_i \in$

$\mathcal{X}$. AXE fits multiple $k$-NN models $M_i^k$ to predict model outputs $Y_{\text{preds}}$, not data labels $Y$. AXE has two hyperparameters: $n$ for the "top-n" number of important features to use and $k$, for the number of neighbors to use in the $k$-NN model $M^k$; denoted $\text{AXE}_n^k$.

It is essential *not* to use the same $k$-NN model for all predictions – AXE does not build a global $k$-NN surrogate to measure explanation quality. For each datapoint $\mathbf{x}_i$, we use a unique $k$-NN model that considers the top-n most important features for that particular explanation $\mathbf{e}_i$. This insight is critical to ensure that AXE does not just report the accuracy of an arbitrary $k$-NN model over the entire dataset, and properly reflects the desiderata from the *local contextualization* principle. By using AXE to predict $Y_{\text{preds}}$ instead of $Y$, we satisfy the *model relativism*, and by using the same training data as $m$, we satisfy *on-manifold evaluation*.

For efficiency, the $k$-NN models $M_i^k$ trained in line 5 can be cached and reused. For a dataset with $v$ features, the number of unique $k$-NN models trained to compute $\text{AXE}_k^n$ is at most $\binom{N}{n}$. The cache size is bound by $\min(v, \binom{N}{n})$. Lastly, AXE is flexible to allow the use of different performance measures in line 9. In our experiments we use accuracy: $\text{Accuracy}(\hat{Y}, Y_{\text{preds}}) = \frac{1}{v}\sum_{i=1}^{v} \mathbf{1}_{(\hat{y}_i = y_i)}$

Instead of selecting a specific value of the hyperparameter $n$, algorithm 1 can be repeated for all possible values $n \in (0, N]$, computing multiple $\text{AXE}_k^n$ scores. Finally, the area under a $n$ – $\text{AXE}_n^k$ curve (AUC) can be used to obtain a single number as an overall evaluation of the model explanation quality, independent of $n$. This AUC trick is adopted from the literature and is a common way to obtain scalar scores from top-n based evaluation metrics [2, 24, 35, 38]. This way, AXE can be made sensitive to the entire order of feature importances instead of just the top-n.

AXE satisfies the *local contextualization* principle by using a different set of neighbors for each datapoint $\mathbf{x}$. Each datapoint has its own nearest-neighbors model $M_i^k$, unique to each prediction $\hat{y}_i$ for $\mathbf{x}_i$. It satisfies the *model relativism* principle by training model $M_i^k$ to predict the classifier response $Y_{\text{preds}}$, rather than the target feature $Y$, making the quality metric $q$ dependent on the model $m$. Finally, AXE satisfies the *on-manifold evaluation* principle because the $k$-NN models are explicitly limited to the existing data manifold and do not rely on new datapoints $\mathbf{x} \notin \mathcal{X}$, avoiding feature sensitivity measures. We can use AXE to determine the quality scores for the explanations in figure 1. Using $\text{AXE}_{n=4}^{k=5}$, we get:- (a) Gradients: 0.4; (b) SHAP: 1.0; (c) LIME: 0.6; and (d) Integrated Gradients: 0.8, indicating SHAP is the most useful for predicting model behavior.

## 3.2 Illustrative Example

Intuitively, AXE uses $k$-NN models because we want important features to be those that separate model predictions in feature space. Conversely, unimportant inputs should be unable to separate the model predictions from each other. To illustrate the intuition behind the choice of $k$-NN models in AXE , we present a motivating example. Consider a dataset consisting of input features $X_1$ and $X_2$, sampled from 4 Normal distributions illustrated in figure 4.

The data (5000 points each) is sampled from 4 Normal distributions $\mathcal{N}_Q$, $\mathcal{N}_R$, $\mathcal{N}_S$, and $\mathcal{N}_T$, with varying covariances, respectively centered at $Q = (2, 2)$, $R = (-2, 2)$, $S = (-2, -2)$, and $T = (2, -2)$. Model $m$ makes predictions using only input feature $X_1$, independent of feature $X_2$. By construction, for all non-outliers we can assume: $m(\mathbf{x}_q) = 1 \forall \mathbf{x}_q \sim \mathcal{N}_Q$, $m(\mathbf{x}_r) = 0 \forall \mathbf{x}_r \sim \mathcal{N}_R$, $m(\mathbf{x}_s) = 0 \forall \mathbf{x}_s \sim \mathcal{N}_S$, and $m(\mathbf{x}_t) = 1 \forall \mathbf{x}_t \sim \mathcal{N}_T$. Explanation $\mathbf{e}_a = [i_{a1}, i_{a2}]$ has $i_{a1} > i_{a2}$ ($X_1$ is more important), and explanation $\mathbf{e}_b = [i_{b1}, i_{b2}]$ has $i_{b1} < i_{b2}$ ($X_2$ is more important). Ideally, metric $q$ should correctly assign a higher score to an explanation $\mathbf{e}_a$: $q(\mathbf{e}_a) > q(\mathbf{e}_b)$. For $\mathbf{e}_a$ (and $\mathbf{e}_b$), the PGI perturbation neighborhood is $\Delta_a$ (and $\Delta_b$) and the AXE $k$-NN neighborhood is $\eta_a$ (and $\eta_b$) along the respective axes $X_1$ (and $X_2$) respectively. In practice, the $\Delta$ and $\eta$ neighborhoods are very similar.

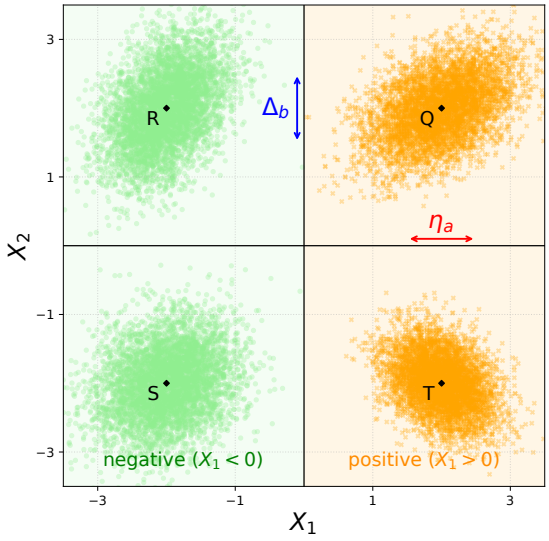
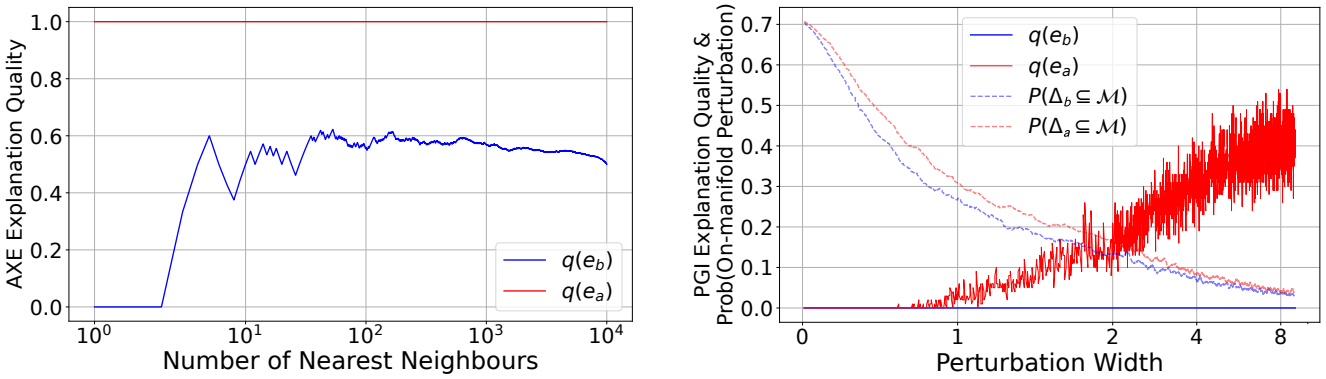



**Figure 4: Synthetic Data and Model for AXE and PGI evaluations: 4 Normal distributions representing the data distribution, and neighborhoods $\Delta$ and $\eta$ for PGI and AXE respectively. The model is defined as $m(\mathbf{x}) = \mathbf{1}_{X_1 > 0}$, and we compare the quality of competing explanations $\mathbf{e}_a$ ($X_1$ is more important) and $\mathbf{e}_b$ ($X_2$ is more important) for datapoint $Q$.**

We restrict our illustrative analysis to explanations for a single datapoint $Q = (2, 2)$. From the definition of model $m$, it is clear that $X_1$ should be more important that $X_2$, and consequently $\mathbf{e}_a$ better than $\mathbf{e}_b$. Upon examination, this can easily be verified to be the case for $\text{AXE}_{n=1}$. Consider $\mathbf{e}_b$ where $X_2$ is the more important feature. A nearest neighbor model finding neighbors for $Q$ considering only $X_2$ and ignoring $X_1$ (per the AXE definition) will find datapoints in the $k$-NN neighborhood $\eta_b$ for point $Q$ and feature $X_2$. This would include points $\mathbf{x}_q$, and $\mathbf{x}_r$, but not $\mathbf{x}_s$ or $\mathbf{x}_t$. Predictions from these points can be either 1 or 0 respectively, hence the nearest neighbor model will have both labels in its neighbors, predicting an average near 0.5. This implies poor accuracy in recovering the positive prediction $m(Q)$, leading to a low AXE score ($\sim 0.5$, see figure 5 a). On the other hand, for explanation $\mathbf{e}_a$, since $X_1$ is the important feature, the neighborhood $\eta_a$ will include points from $\mathbf{x}_q$ and $\mathbf{x}_t$. These are all predicted to fall in the positive class, thus recovering $m(Q)$ with perfect accuracy of 1.0 and leading to a high AXE score ($\sim 1.0$, see figure 5 a). Figure 5 (a) plots $\text{AXE}_{n=1}(\mathbf{e}_a)$ and $\text{AXE}_{n=1}(\mathbf{e}_b)$ for different $k$ values for the $k$-NN models in AXE, clearly showing $\text{AXE}_n(\mathbf{e}_a) > \text{AXE}_n(\mathbf{e}_b)$ for all hyperparameter values of $k$.




**(a) AXE: AXE relaibly shows that explanation $\mathbf{e}_a$ is better than explanation $\mathbf{e}_b$. AXE $_{n=1}(\mathbf{e}_a) >$ AXE $_{n=1}(\mathbf{e}_b) \forall k \in (1, 10000)$.**

**(b) PGI: $\text{PGI}_{n=1}(\mathbf{e}_a) > \text{PGI}_{n=1}(\mathbf{e}_b)$ only once the perturbations are large, where on-manifold probability $P(\Delta \subseteq \mathcal{M})$ is low.**

**Figure 5: Comparing explanations using AXE and PGI: By definition $q(\mathbf{e}_a) > q(\mathbf{e}_b)$, but PGI does not clearly show this. AXE correctly determines that explanation $\mathbf{e}_a$ is better than $\mathbf{e}_b$, across hyperparameter values. (Both X axes on symlog scale).**

We now analyze the behavior of PGI. For $\mathbf{e}_b$, PGI would generate datapoints $(2.0, 2.0+\delta), \delta \sim \mathcal{N}(0, \text{width})$. Varying $X_2$ has no impact on the model prediction (by definition), yielding a prediction gap of 0. Conversely, for $\mathbf{e}_a$, PGI would sample datapoints $(2.0+\delta, 2.0), \delta \sim \mathcal{N}(0, \text{width})$. The predictions for these are highly sensitive to the neighborhood $\Delta_a$, the the PGI sampling width. This hyperparameter determines whether the neighborhood stays on the same side of the decision boundary $X_1 = 0$. If it does, then PGI is 0 – a result that provides no information to compare $\mathbf{e}_a$ and $\mathbf{e}_b$. Figure 5 (b) shows that PGI is zero until the neighborhood becomes large enough. However large neighborhoods present a different challenge – the points PGI samples are more likely to lie off manifold. Figure 5 (b) also shows the corresponding probability of the PGI perturbations lying on manifold, and it can be seen that in the "useful" non-zero range of the plot, on-manifold probabilities are lower. In general, it is difficult to tune the neighborhood width hyperparameter in sensitivity analysis [40, 49]. Tuning this hyperparameter requires knowing apriori what explanations to expect – an implausible expectation akin to knowing "ground-truth" explanations. Lastly, figure 5 (b) shows that PGI is unstable even in regions of high perturbation width, further complicating its use in practice.

AXE does not require model predictions on off-manifold data. AXE also does not need access to ground-truth explanations. This is enabled by using $k$-NN model accuracy to measure explanation quality – fitting a different $k$-NN model for each unique explanation and datapoint. Critically, using $k$-NN models provides a number of advantages. $k$-NN models operate on the classifier data itself, omitting the need for off-manifold predictions, satisfying the *on-manifold evaluation* principle. Further, the choice of nearest neighbor models directly captures the notion of separability in feature space – capturing the idea that an important feature is one that separates classes in feature space.

## 4 Experiments

In sections 2.4, 2.5, and 3.2 we used synthetic examples demonstrating the ineffectiveness of prior explanation evaluation metrics. We now compare AXE with existing baselines on real-world datasets.

### 4.1 Detecting Explanation Fairwashing

We simulate a state-of-the-art adversarial attack [54] on explanations in a real-world setting where ground-truths remain unknown. The attack modifies a model $m$, known to be discriminatory, creating new models $m_S$ or $m_L$ that respectively fool SHAP and LIME into generating explanations $\mathbf{e}$ that show the discriminatory feature as unimportant. A good explanation evaluation metric $q$ should identify manipulated explanations by scoring them poorly.

This attack fairwashes model explanations by hiding discriminatory model behavior. Imagine explanations for the diabetes prediction model figure 1 showing that the model used benign inputs to make its prediction, when the model actually made predictions using protected attributes that are medically irrelevant.

Like $m$, $m_S$ and $m_L$ too make decisions using only the "protected" feature ($X_\rho$), but they fool explainers SHAP and LIME respectively into generating explanations showing spurious "foil" features ($X_\phi, X_\psi$) as the most important [54]. We use the same datasets and models as the original adversarial attack [54]: the German Credit dataset from lending [27] and the COMPAS [32] dataset and the Communities and Crime dataset [47] from criminal justice.

We then manually construct competing feature-importance explanations: $E_\rho$ is the set of all explanations where the protected feature $X_\rho$ is the most important feature. $E_\phi$ and $E_\psi$ are the explanation sets where the foil feature $X_\phi$ or $X_\psi$ is the most important, respectively. Finally, $E_\omega$ is the set of all other explanations, where the most important feature is neither protected nor a foil feature. The credit model is adversarially modified to deliberately favor men over others, and the criminal justice models are modified to favor white people over others. Explanations that reveal this ($E_\rho$) are correct. Explanations that mask this ($E_\phi, E_\psi$) are spurious. Evaluation metrics $q$ should identify this by the condition $\bar{q}(E_\rho) > \bar{q}(E_\phi)$ and $\bar{q}(E_\rho) > \bar{q}(E_\psi)$; where $\bar{q}$ is the average quality $\bar{q}(E) = \sum_{\mathbf{e} \in E} q(\mathbf{e})/|E|$.

In table 2 we summarize our results. We use AXE $_{n=1}$, $\text{PGI}_{n=1}$ and $\text{PGU}_{n=1}$ to measure the quality of the 3 sets of explanations. We fix top-n as 1 in each case because our explanations are constructed

**Table 2: Detecting explanation fairwashing: We replicate an adversarial fairwashing attack [54] and generate spurious explanations, which we try to then detect using quality metrics $q$ that do not need ground-truths: AXE , PGI, and PGU. Complete details are in section 4.1: $E_\rho$ is a set of explanations that correctly denote that the most important model input feature is $X_\rho$. $E_\phi$ and $E_\psi$ are sets of manipulated explanations created by an adversary where the most important feature is $X_\phi$ or $X_\psi$.**
**A good evaluation metric $q$ should distinguish manipulated explanations from correct explanations – we expect that: (i) $\bar{q}(E_\rho) > \bar{q}(E_\phi)$ and (ii) $\bar{q}(E_\rho) > \bar{q}(E_\psi)$; where $\bar{q}(E) = \sum_{\mathbf{e}\in E} q(\mathbf{e})/|E|$. Cases where only one of the two conditions is a strict inequality are marked with an asterisk*. AXE has a success rate of 100%, whereas the overall success rate for PGI and PGU is only 50%.**

| Dataset | Adversarial Model $m_L$ or $m_S$ | Eval. Metric $q$ ($n=1$) | *Evaluating explanations with a single important attribute:* Protected $q(E_\rho)$ | Foil 1 $q(E_\phi)$ | Foil 2 $q(E_\psi)$ | Other $q(E_\omega)$ | $q(E_\rho) > q(E_\phi)$ *and* $q(E_\rho) > q(E_\psi)$ |
|---|---|---|---|---|---|---|---|
| German Credit | $m_L$ *(1 foil)* | PGI | 0.032 | 0.148 | na | 0.018 | ✗ |
| | | (-)PGU | -0.486 | -0.536 | na | -0.483 | ✓ |
| | | AXE | 1.000 | 0.680 | na | 0.617 | ✓ |
| | $m_S$ *(1 foil)* | PGI | 0.037 | 0 | na | 0.037 | ✓ |
| | | (-)PGU | -0.475 | -0.529 | na | -0.478 | ✓ |
| | | AXE | 0.990 | 0.690 | na | 0.622 | ✓ |
| COMPAS | $m_L$ *(1 foil)* | PGI | 0.006 | 0 | na | 0.067 | ✓ |
| | | (-)PGU | -0.481 | -0.479 | na | -0.431 | ✗ |
| | | AXE | 0.992 | 0.739 | na | 0.534 | ✓ |
| | $m_S$ *(1 foil)* | PGI | 0.006 | 0.035 | na | 0.009 | ✗ |
| | | (-)PGU | -0.091 | -0.077 | na | -0.090 | ✗ |
| | | AXE | 0.968 | 0.761 | na | 0.527 | ✓ |
| | $m_L$ *(2 foils)* | PGI | 0.006 | 0 | 0.001 | 0.075 | ✓ |
| | | (-)PGU | -0.520 | -0.520 | 0-0.524 | -0.464 | ✗* |
| | | AXE | 0.990 | 0.739 | 0.735 | 0.533 | ✓ |
| | $m_S$ *(2 foils)* | PGI | 0.005 | 0.039 | 0.041 | 0.010 | ✗ |
| | | (-)PGU | -0.104 | -0.090 | -0.092 | -0.106 | ✗ |
| | | AXE | 0.956 | 0.746 | 0.731 | 0.531 | ✓ |
| Communities and Crime | $m_L$ *(1 foil)* | PGI | 0.103 | 0 | na | 0.029 | ✓ |
| | | (-)PGU | -0.479 | -0.460 | na | -0.481 | ✗ |
| | | AXE | 1.000 | 0.765 | na | 0.793 | ✓ |
| | $m_S$ *(1 foil)* | PGI | 0.089 | 0.006 | na | 0.005 | ✓ |
| | | (-)PGU | -0.446 | -0.429 | na | -0.448 | ✗ |
| | | AXE | 0.985 | 0.765 | na | 0.790 | ✓ |
| | $m_L$ *(2 foils)* | PGI | 0.101 | 0.001 | 0.001 | 0.034 | ✓ |
| | | (-)PGU | -0.534 | -0.536 | -0.536 | -0.535 | ✓ |
| | | AXE | 0.995 | 0.760 | 0.760 | 0.792 | ✓ |
| | $m_S$ *(2 foils)* | PGI | 0.094 | 0.006 | 0.005 | 0.008 | ✓ |
| | | (-)PGU | -0.479 | -0.470 | -0.479 | -0.479 | ✗* |
| | | AXE | 0.955 | 0.760 | 0.755 | 0.781 | ✓ |

to only promote one feature as important at a time. As explained, our verification test for AXE is that $\text{AXE}(E_\rho) > \text{AXE}(E_\phi)$ and $\text{AXE}(E_\rho) > \text{AXE}(E_\psi)$, which is found to always be true. Both PGI and PGU fail their corresponding checks.

The last column of table 2 shows PGU failing to discern genuine explanations $E_\rho$ from spurious ones $E_\psi$, $E_\phi$ 7 out of 10 times, and PGI failing to do so 3 out of 10 times. AXE never fails, placing the overall error rate for sensitivity metrics PGI and PGU at 50%, and for AXE at 0%. This indicates the evaluation metrics PGI and PGU are not impartial – their optimization objectives are so aligned with LIME and SHAP that adversarial models designed to fool LIME and SHAP end up fooling PGI and PGU too. Table 2 proves experimentally that PGI and PGU violate the on-manifold evaluation principle, as we showed theoretically in section 2.5.

## 4.2 Comparing AXE with Baselines

We performed computational experiments comparing AXE with prior baselines across four datasets, two models, seven explainers, and all eight prior evaluation metrics from table 1, adopting the standard OpenXAI benchmark [2], with results presented in figures 6 and 7. Like OpenXAI, we use the German Credit [27], COMPAS [32], Adult Income [6], and Home Equity Line of Credit (HELOC) [18] datasets ($\mathcal{X}$), and run experiments with both linear regression and neural network models ($m$). We try the SHAP and LIME (perturbation based); SmoothGrad, Grad, Input × Grad and Integrated Gradients (gradient based); and Random explainers ($\mathcal{E}$). We report results using the FA, RA, SA, SRA, RC, PRA, PGI, and PGU evaluation metrics ($q$).

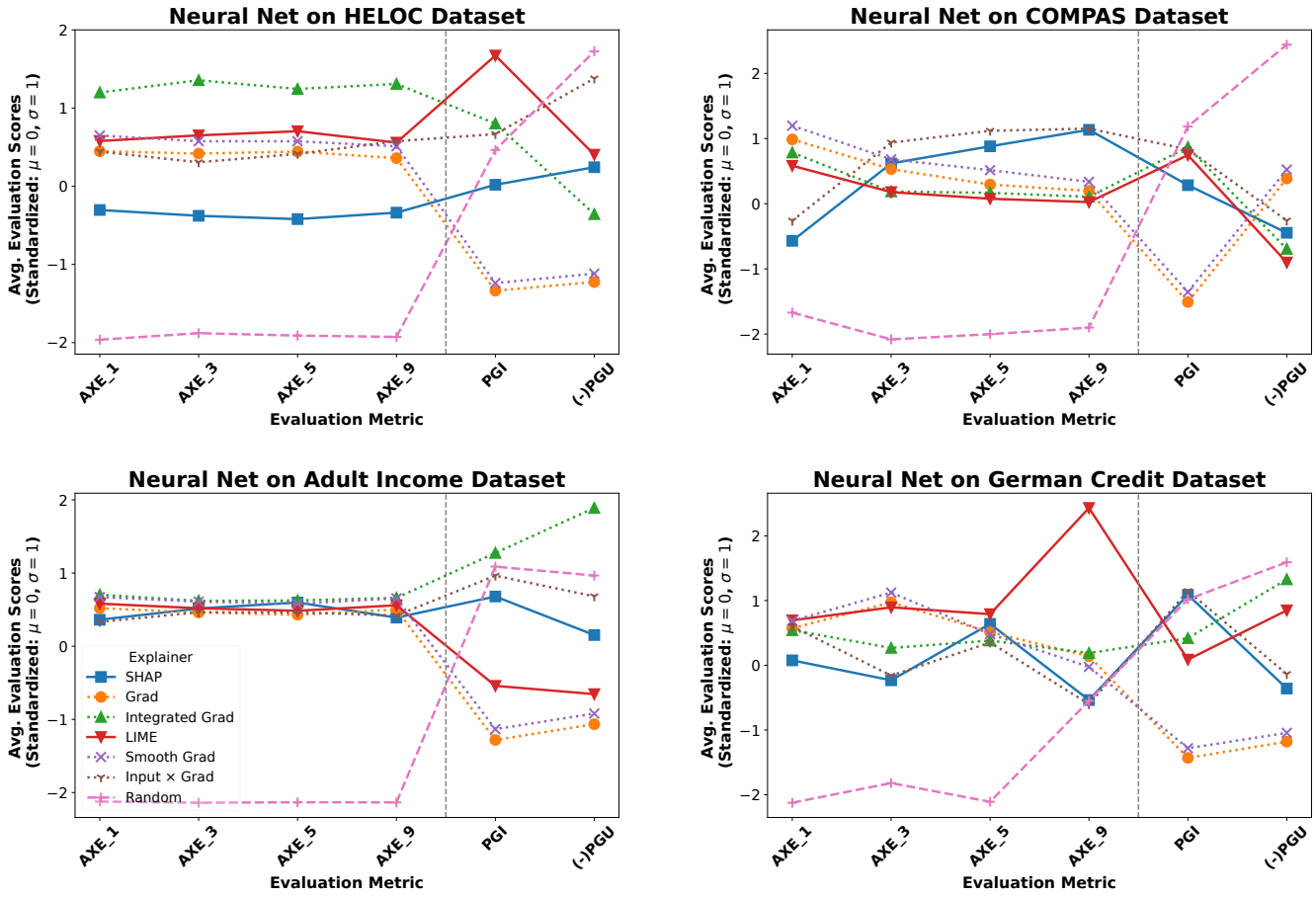



**Figure 6: Evaluating the Quality of Explanations for Neural Networks: For a fair comparison across evaluation metrics, we average over the entire dataset and plot their Z-score standardized values. Neural nets have no ground truth explanation $e^*$, so the only metrics available are AXE, PGI, and -PGU (PGU inverted so higher values are better). Instead of a particular number of top-n features, we use the AUC trick from section 3.1. For details see section 4.2.**

We compare these baselines with $AXE_n^k$. Instead of selecting a particular number of top-n features, we use the AUC trick described in section 3.1 for all evaluation metrics $q$. Instead of selecting a particular value for the $k$-NN hyperparameter $k$, we report results for several values: $AXE^1$, $AXE^3$, $AXE^5$, and $AXE^9$. Each explainer type (perturbation, gradient, or random) is denoted with a different line-style. To compare evaluation metrics with each other, we standardize the final results for each explainer and evaluation metric using z-scores, because they may follow different scales. For instance, while ideal explanations for both AXE and PGI have scores of 1.0, AXE considers uninformative explanations to have values near 0.5, whereas PGI considers uninformative values to be near 0. Additionally, we also invert PGU values (denoted as (-)PGU) so that higher values are better, like the rest of our metrics.

For logistic regression models, we are able to use the ground-truth evaluation metrics because of the presence of model coefficients, which we adopt as ground-truth for every datapoint in the dataset, following previous benchmarks [2, 31]. This approach is discussed in detail in section 2.4. For neural network models, we are only able to use sensitivity based metrics PGI and PGU, because of the lack of ground truth explanations. The results from the logistic regression comparisons can be seen in figure 7 and from the neural network in figure 6.

From the plots in figures 6 and 7, the $AXE^1$, $AXE^3$, $AXE^5$, and $AXE^9$ metrics can be seen to broadly agree with each other in score, further reinforcing the intuition from section 3.2 that AXE is fairly robust to hyperparameter variations. The ground-truth oriented metrics (FA, SA, RA, SRA, RC, PRA) show significant disagreement with each other, as has been noted in the literature [5, 31, 39].

Finally, as a simple check for evaluation metric validity, we focus on the behavior of the Random explainer. Ideally, a good evaluation framework would clearly and reliably distinguish this explainer from the others, however this does not seem to be the case for any previous evaluation frameworks. The sensitivity oriented metrics (PGI and PGU) rank the Random explainer particularly well. This is expected from prior work [49] and from our analysis from section 3.2 where we uncovered the dependence of PGI values on

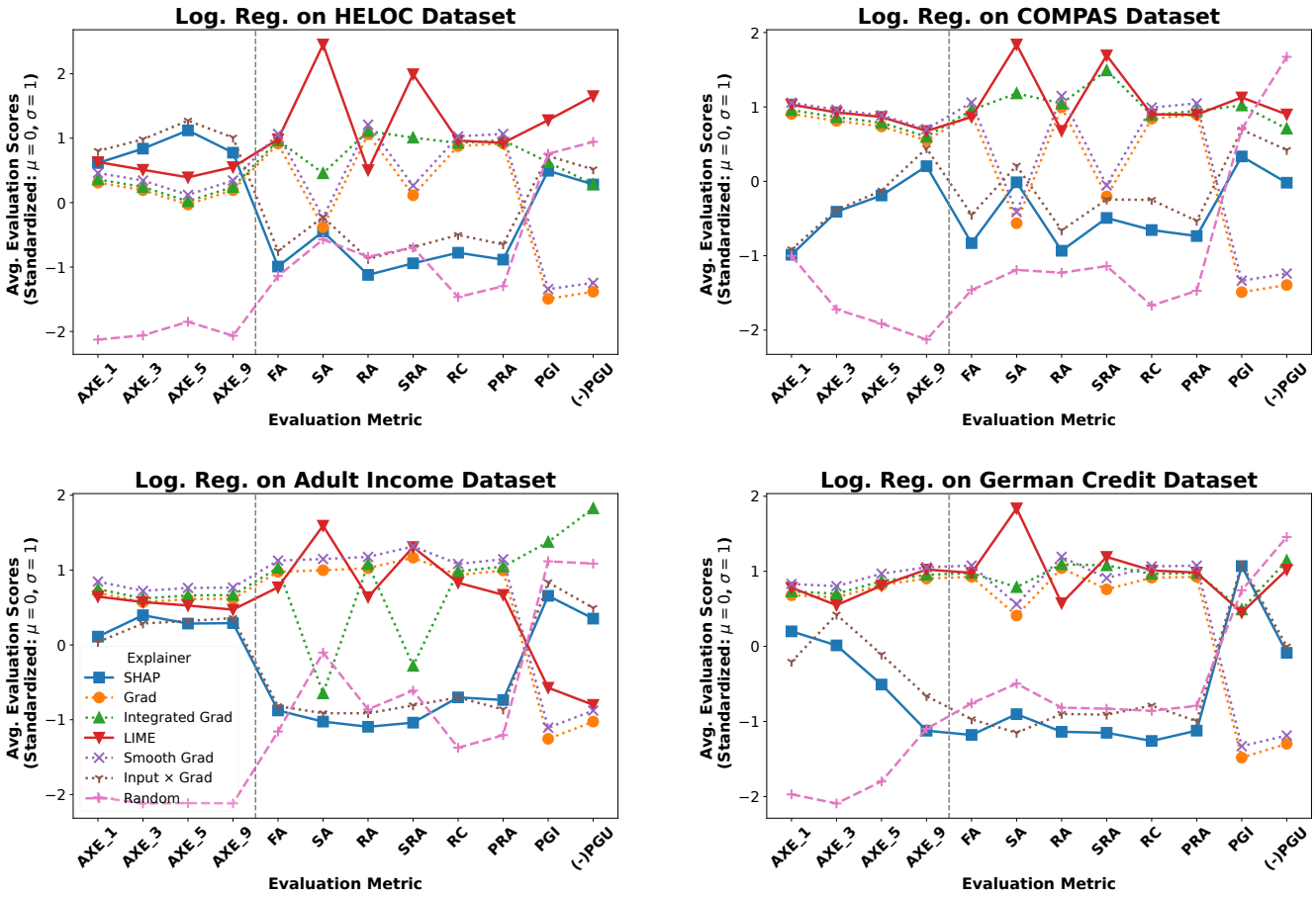



**Figure 7: Evaluating the Quality of Explanations for Logistic Regression: For a fair comparison across evaluation metrics, we average over the entire dataset and plot their Z-score standardized values. We compare AXE with all prior metrics from table 1: FA, RA, SA, SRA, RC, PRA, PGI, and -PGU (PGU inverted so higher values are better). Instead of a particular number of top-n features, we use the AUC trick from section 3.1. For details see section 4.2.**

the neighborhood perturbation width hyperparameter. This further questions the use of feature sensitivity as an effective strategy to evaluate explanation quality, bolstering the importance of the *on-manifold evaluation* principle.

## 5 Conclusion

The ability to measure explanation quality is integral to developing better explanations, and to engender trust in existing XAI methods.

In section 2.2 we introduced three principles to guide the evaluation of explainers and local feature-importance model explanations: **local contextualization**, **model relativism**, and **on-manifold evaluation**, reflecting that explanations should be dependent on input datapoints, dependent on models, and independent of off-manifold behavior respectively. We constructed simple examples in sections 2.4 and 2.5 showcasing the violations of these principles by prior evaluation frameworks, and uncovering the absurdities of existing XAI evaluation frameworks – such as comparing a single global explanation with local model explanations of different input datapoints. To operationalize the finding from human-centered user research that useful explanations are those that help users predict model behavior [12], in section 3.1 we proposed AXE : a new ground-truth **Agnostic eXplanation Evaluation** framework, and in 3.2 we used a simple example to showcase AXE in action and motivate the underlying design choice of $k$-NN. Finally, in section 4.1 we showed empirically how AXE can be used to detect fairwashing of explanations – to our knowledge the first evaluation metric to be able to do this perfectly, and in section 4.2 we compared AXE with prior baselines to show through computations that AXE satisfies all three desirable principles of explanation evaluation.

This work has several implications for AI trustworthiness, fairness, and transparency. The lack of good selection processes to choose between explanations undermines trust not just in individual explanations and models, but in the field at large. It hinders practitioners from adopting XAI, and leads to unresolved problems about explanation disagreement in machine learning. We hope the principles introduced in this paper and the AXE evaluation framework can help build a robust and stable foundation for local explanations in XAI.

## Acknowledgments

This work has been supported through research funding provided by the Wellcome Trust (grant no. 223765/Z/21/Z), Sloan Foundation (grant no. G-2021-16779), Department of Health and Social Care, EPSRC (grant no. EP/Y019393/1), and Luminate Group. Their funding supports the Trustworthiness Auditing for AI project and the Governance of Emerging Technologies research programme at the Oxford Internet Institute, University of Oxford. The donors had no role in the decision to publish or the preparation of this paper.

## References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 9525–9536.

[2] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2024. OpenXAI: towards a transparent evaluation of post hoc model explanations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1148, 16 pages.

[3] Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. 2024. Characterizing the risk of fairwashing. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1136, 13 pages.

[4] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 275–285. doi:10.1145/3377325.3377519

[5] Brian Barr, Noah Fatsi, Leif Hancox-Li, Peter Richter, Daniel Proano, and Caleb Mok. 2023. The Disagreement Problem in Faithfulness Metrics. arXiv:2311.07763 [cs.LG] https://arxiv.org/abs/2311.07763

[6] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[7] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. doi:10.1145/3377325.3377498

[8] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019). doi:10.3390/electronics8080832

[9] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 1036–1042. doi:10.18653/v1/D18-1128

[10] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. 2022. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput Appl* , (March 2022), 1–17.

[11] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1. doi:10.1145/3593013.3593970

[12] Julien Colin, Thomas FEL, Remi Cadene, and Thomas Serre. 2022. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., , 2832–2845. https://proceedings.neurips.cc/paper_files/paper/2022/file/13113e938f2957891c0c5e8df811dd01-Paper-Conference.pdf

[13] T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27. doi:10.1109/TIT.1967.1053964

[14] Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H. Bach, and Himabindu Lakkaraju. 2022. Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 203–214. doi:10.1145/3514094.3534159

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

[16] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML] https://arxiv.org/abs/1702.08608

[17] Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. 2024. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1991, 10 pages.

[18] FICO. 2022. Explainable Machine Learning Challenge. https://community.fico.com/s/explainable-machine-learning-challenge?tabset-158d9=3.

[19] Evelyn Fix and J. L. Hodges. 1989. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique* 57, 3 (1989), 238–247. http://www.jstor.org/stable/1403797

[20] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. , , 2950–2958. doi:10.1109/ICCV.2019.00304

[21] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. , , 3449–3457. doi:10.1109/ICCV.2017.371

[22] Benjamin Fresz, Lena Lörcher, and Marco Huber. 2024. Classification Metrics for Image Explanations: Towards Building Reliable XAI-Evaluations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3630106.3658537

[23] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. , , 80–89. doi:10.1109/DSAA.2018.00018

[24] Thomas Hartley, Kirill Sidorov, Christopher Willis, and David Marshall. 2020. Explaining Failure: Investigation of Surprise and Expectation in CNNs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. , , 56–65. doi:10.1109/CVPRW50498.2020.00014

[25] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5540–5552. doi:10.18653/v1/2020.acl-main.491

[26] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* , (2021), .

[27] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[28] Sara Hooker, Dumitru Erhan, Pieter jan Kindermans, and Been Kim. 2018. Evaluating Feature Importance Estimates. *arXiv* , (2018), . https://arxiv.org/pdf/1806.10758.pdf

[29] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. 2021. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recogn. Lett.* 150, C (Oct. 2021), 228–234. doi:10.1016/j.patrec.2021.06.030

[30] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. 2019. XRAI: Better Attributions Through Regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. , , 4947–4956. doi:10.1109/ICCV.2019.00505

[31] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2024. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *Transactions on Machine Learning Research* , (2024), . https://openreview.net/forum?id=jESY2WTZCe

[32] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[33] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *CoRR* abs/2110.10790 (2021). arXiv:2110.10790 https://arxiv.org/abs/2110.10790

[34] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (June 2018), 31–57. doi:10.1145/3236386.3241340

[35] Yang Liu, Sujay Khandagale, Sujay Khandagale, Colin White, and Willie Neiswanger. 2021. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. , . https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c16a5320fa475530d9583c34fd356ef5-Paper-round2.pdf

[36] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural*

*Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[37] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, , 6765–6774. https://proceedings.mlr.press/v119/marx20a.html

[38] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. 2022. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports* 12, 1 (03 May 2022), 7166. doi:10.1038/s41598-022-11012-2

[39] Vishwali Mhasawade, Salman Rahman, Zoé Haskell-Craig, and Rumi Chunara. 2024. Understanding Disparities in Post Hoc Machine Learning Explanation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2374–2388. doi:10.1145/3630106.3659043

[40] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). leanpub, . https://christophm.github.io/interpretable-ml-book

[41] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In *ECML PKDD 2020 Workshops*, Irena Koprinska, Michael Kamp, Annalisa Appice, Corrado Loglisci, Luiza Antonie, Albrecht Zimmermann, Riccardo Guidotti, Özlem Özgöbek, Rita P. Ribeiro, Ricard Gavaldà, João Gama, Linara Adilova, Yamuna Krishnamurthy, Pedro M. Ferreira, Donato Malerba, Ibéria Medeiros, Michelangelo Ceci, Giuseppe Manco, Elio Masciari, Zbigniew W. Ras, Peter Christen, Eirini Ntoutsi, Erich Schubert, Arthur Zimek, Anna Monreale, Przemyslaw Biecek, Salvatore Rinzivillo, Benjamin Kille, Andreas Lommatzsch, and Jon Atle Gulla (Eds.). Springer International Publishing, Cham, 417–431.

[42] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55, 13s, Article 295 (July 2023), 42 pages. doi:10.1145/3583558

[43] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 3126–3132. doi:10.1145/3366423.3380087

[44] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. *CoRR* abs/1806.07421 (2018), . arXiv:1806.07421 http://arxiv.org/abs/1806.07421

[45] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. doi:10.1145/3411764.3445315

[46] Luyu Qiu, Yi Yang, Caleb Chen Cao, Yueyuan Zheng, Hilary Ngai, Janet Hsiao, and Lei Chen. 2022. Generating Perturbation-based Explanations with Robustness to Out-of-Distribution Data. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3594–3605. doi:10.1145/3485447.3512254

[47] Michael Redmond. 2002. Communities and Crime. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53W3X.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[49] Richard D Riley and Gary S Collins. 2023. Stability of clinical prediction models developed using statistical or machine learning methods. *Biom J* 65, 8 (July 2023), e2200302.

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

[51] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. , , 618–626. doi:10.1109/ICCV.2017.74

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

[53] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). , , . http://arxiv.org/abs/1312.6034

[54] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2019. How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods. *CoRR* abs/1911.02508 (2019), . arXiv:1911.02508 http://arxiv.org/abs/1911.02508

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

[56] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, , 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

[57] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-Down Neural Attention by Excitation Backprop. *Int. J. Comput. Vision* 126, 10 (oct 2018), 1084–1102. doi:10.1007/s11263-017-1059-x

[58] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* 10, 5 (2021), . doi:10.3390/electronics10050593