Evaluating the Ability of Explanations to Disambiguate Models in a Rashomon Set

Kaivalya Rawal[1], Eoin Delaney[2], Zihao Fu[3], Sandra Wachter[1,4], Chris Russell[1]
[1]Oxford Internet Institute, University of Oxford
[2]Trinity College Dublin
[3]Chinese University of Hong Kong
[4]Hasso Plattner Institute
kaivalya.rawal@oii.ox.ac.uk, eoin.delaney@tcd.ie, zihaofu@cuhk.edu.hk, sandra.wachter@hpi.de, chris.russell@oii.ox.ac.uk

Abstract

Explainable artificial intelligence (XAI) is concerned with producing explanations indicating the inner workings of models. For a Rashomon set of similarly performing models, explanations provide a way of disambiguating the behavior of individual models, helping select models for deployment. However explanations themselves can vary depending on the explainer used, and need to be evaluated. In the paper "Evaluating Model Explanations without Ground Truth", we proposed three principles of explanation evaluation and a new method "AXE" to evaluate the quality of feature-importance explanations. We go on to illustrate how evaluation metrics that rely on comparing model explanations against ideal ground truth explanations obscure behavioral differences within a Rashomon set. Explanation evaluation aligned with our proposed principles would highlight these differences instead, helping select models from the Rashomon set. The selection of alternate models from the Rashomon set can maintain identical predictions but mislead explainers into generating false explanations, and mislead evaluation methods into considering the false explanations to be of high quality. AXE, our proposed explanation evaluation method, can detect this adversarial fairwashing of explanations with a 100% success rate. Unlike prior explanation evaluation strategies such as those based on model sensitivity or ground truth comparison, AXE can determine when protected attributes are used to make predictions.

Previously Reviewed Extended Version —
https://dl.acm.org/doi/10.1145/3715275.3732219

Code and Datasets —
https://github.com/KaiRawal/Evaluating-Model-Explanations-without-Ground-Truth

Introduction

Models that make similar predictions on a given domain of input datapoints form a Rashomon set of models. These models can differ in their internal decision-making mechanisms, causing differences in critical aspects such as fairness towards protected groups or the use of protected input features. A set of models forming a Rashomon set might have similar predictions in a certain input manifold, but have divergent behavior in a different region of input datapoints. With the increasingly complex and opaque inner workings of artificial intelligence systems, the internal mechanisms underlying models often need to be discovered through dedicated explainable artificial intelligence (XAI) techniques. Explainability techniques can help disambiguate models from a Rashomon set with otherwise indistinguishable outputs, and aid in model selection by operationalizing criteria beyond model accuracy.

While there are many competing explanation types, data modalities, and evaluation desiderata, (Liao and Varshney 2021), this paper focuses exclusively on local feature-importance explanations for models operating on tabular datasets, such as LIME (Ribeiro, Singh, and Guestrin 2016) or SHAP (Lundberg and Lee 2017). These can operate on any model type and explain individual predictions rather than describing global model behavior, producing feature importances as output: a signed vector indicating the relative contribution of each input feature to the model prediction.

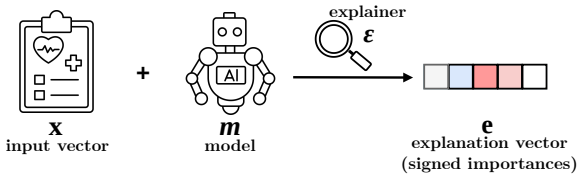
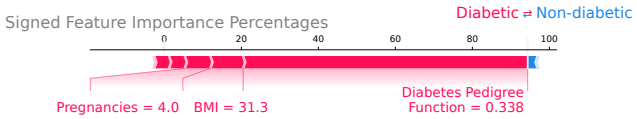


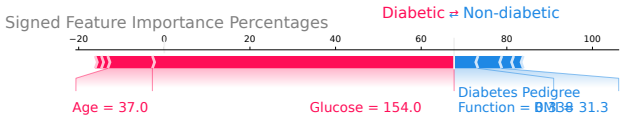
Figure 1: **Explanation generation**: Explainer $\mathcal{E}$ produces an explanation vector $\mathbf{e}$ of signed feature importances using datapoint $\mathbf{x}$, model m and prediction $m(\mathbf{x})$.

Even in this restricted setting, different explanation methods (explainers) often provide contradictory explanations, such as in figure 2. The selection of incorrect or intentionally misleading explanations can misinform users and regulators, reinforcing systemic biases, and eroding public trust in AI systems (Barr et al. 2023; Mhasawade et al. 2024; Marx, Calmon, and Ustun 2020). Evaluating the quality of generated explanations is therefore a critically important problem. This is often done anecdotally – of approximately 300 papers proposing new model explanation methods (explainers), one in three papers evaluated explanations anecdotally (Nauta et al. 2023). Without consensus on the essential prop-

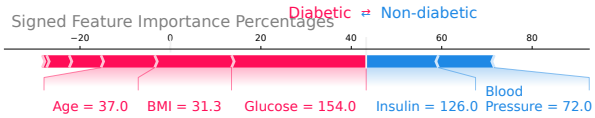
1st Workshop on Navigating Model Uncertainty and the Rashomon Effect: From Theory and Tools to Applications and Impact
Copyright © 2026, Association for the Advancement of Artificial Intelligence (AAAI'26). http://www.aaai.org



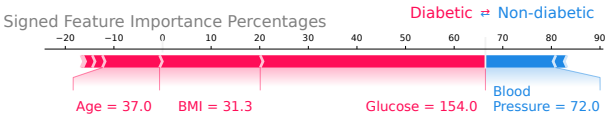
(a) **"Gradients"** explainer: **Diabetes Pedigree Function** is the most important input, pushing the model to a positive prediction (diabetic)



(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 2: **Different Explainers Yield Different Explanations**: A neural network predicts diabetes on the "Pima Indians" dataset (Chang et al. 2022). A single positive (diabetic) prediction is explained using four explainers. Each feature-importance explanation varies, but consists of a signed vector indicating the relative contribution of each input to the model output.

erties that explanations should possess and robust frameworks to numerically evaluate them, progress in the field remains slow, with fragmented benchmarks.

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. When seeking to disambiguate the behavior of competing models from a Rashomon set, explanations should seek to highlight model differences rather than obscure them. This is operationalized through the *model relativism* principle, which promotes explanations being dependent on the specific model used, rather than aligning towards an ideal value common to the entire Rashomon set. Conversely, explanations for models in a Rashomon set should be the same whenever the internal decision making mechanisms are the same. The *on-manifold evaluation* principle requires explanations to depend only on on-manifold model behavior, discounting off-manifold behavior and promoting similar explanations for genuinely similar models.

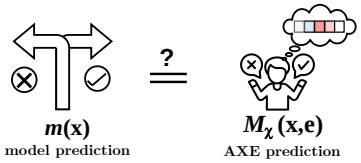


Figure 3: **Explanation evaluation**: AXE evaluates the quality q of explanation $\mathbf{e}$ by measuring how accurately prediction $m(\mathbf{x})$ can be recovered from dataset $\mathcal{X}$.

We use these principles to propose AXE, a new feature sensitivity and 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 (Colin et al. 2022). AXE satisfies the *model relativism* principle, and can be used to distinguish models from a Rashomon set that make different predictions. It also simultaneously satisfies the *on-manifold evaluation* principle, ascribing the same explanation to models in a Rashomon set if they do use the same internal decision mechanisms. Adversarial attacks have been proposed to exploit this latter vulnerability, where a discriminatory model's behavior is masked using another model that mimics discriminatory predictions (thus forming part of a Rashomon set), but varies in off-manifold behavior to mislead explainers into masking its discriminatory behavior. AXE is invulnerable to this attack and is able to identify the correct explanation despite the adversarial manipulation.

Our notation: For ease of exposition we specify our notation from figures 1 and 3: 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.

- *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.
- *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: $m(\mathbf{x}) = y_{\text{pred}} \in \{0, 1\}$.
- *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 $m(\mathbf{x})$.

- *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 (Carvalho, Pereira, and Cardoso 2019; Liu et al. 2021; Gilpin et al. 2018).

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

- $\mathcal{X} \in \mathbb{R}^{\nu \times N}$ is the dataset of inputs with N features and ν datapoints, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\nu\}$.
- $m : \mathbb{R}^N \to \{0, 1\}$ 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:

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

Three Principles for Evaluating Explanations

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 $m_1 \neq m_2$ typically have different explanations; and (c) as seen in figure 2, explanations from different explainers can have different explanations 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 selecting a different model m from the Rashomon set that makes similar predictions but uses a different internal mechanism, the evaluation metric q should promote corresponding changes in the explanation $\mathbf{e}$. Explanations should not hide, but surface behavioral differences across models in a Rashomon set.
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 $\mathbf{e}$. Different models from a Rashomon set have the same behavior on the data manifold, but might vary in the predictions made "off-manifold". Explanations should be similar for such models, and evaluation metrics should consequently be independent of model behavior in "explanation manifolds".

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 (Ivanovs, Kadikis, and Ozols 2021; Qiu et al. 2022; Fel et al. 2024). Ideally, an explanation for model behavior on datapoint $\mathbf{x_1}$ should not depend on model 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

Table 1: **Explanation Evaluation Metrics**: Definitions for ground-truth based explanation evaluation metrics (for explanation $\mathbf{e}$ and ground truth $\mathbf{e}^*$): **FA**, **RA**, **SA**, **SRA**, **RC** and **PRA** (Krishna et al. 2024; Agarwal et al. 2024); sensitivity based metrics **PGI** and **PGU** (Agarwal et al. 2024; Dai et al. 2022; Pawelczyk, Broelemann, and Kasneci 2020); and AXE. We list whether the metric satisfies each principle: Local Contextualization (**P1**), Model Relativism (**P2**), and On-Manifold Evaluation (**P3**).

Metric	Definition	P1	P2	P3
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 algorithm 1.**	✔	✔	✔

neighborhoods around real points, are simply encoding a particular choice of sensitivity analysis without meaningfully evaluating the explanation quality.

Broadly, prior explanation evaluation metrics can be classified into two types:

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 (Petsiuk, Das, and Saenko 2018; Fong and Vedaldi 2017; Fong, Patrick, and Vedaldi 2019; Selvaraju et al. 2017). These include Feature Agreement (FA), Sign Agreement (SA), Rank Agreement (RA), Signed Rank Agreement (SRA), Rank Correlation (RC) and Pairwise Rank Agreement (PRA) (Krishna et al. 2024).
2. **Sensitivity** based metrics verify model sensitivity to the inputs declared important by an explanation (Petsiuk, Das, and Saenko 2018; Fel et al. 2024; Samek et al. 2017; Kapishnikov et al. 2019). These have been summarized as Prediction Gap on Important Feature Perturbation (PGI) and Prediction Gap on Unimportant Feature Perturbation (PGU) (Dai et al. 2022; Pawelczyk, Broelemann, and Kasneci 2020; Agarwal et al. 2024).

Sensitivity based metrics overindex on particular explanation methods. For instance, by defining the perturbation method appropriately, PGI can be made to emulate SHAP or LIME exactly, thus scoring particular explainers perfectly. For ground-truth based metrics, there can be similar variations in the notion of "ground-truth". Proposals include one "ground-truth" per datapoint $\mathbf{x}$, unintentionally introducing independence from m (Zhang et al. 2018; Adebayo et al. 2018) violating the *model relativism* principle. Consider a datapoint x such that all models in a Rashomon set $m_i \in \mathcal{M}$ produce an identical prediction. By having an immutable "ground-truth" $\mathbf{e}^*$ to compare diverse explanations e with, the evaluation metric incentivizes explainers to output the same explanation regardless of the particular model m_i.

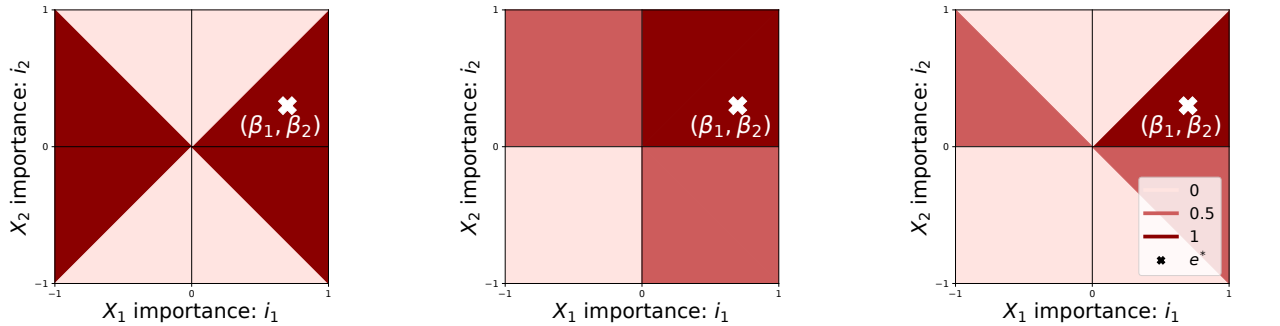
With images especially, an explanation is often considered good if it selects the "correct" region as important in an image – regardless of whether the model used those features (Zhang et al. 2018; Adebayo et al. 2018; Fresz, Lörcher, and Huber 2024). This leads to egregious violations of the model relativism principle, as many models m_i detecting the same object in an image (but due to different pixels) can form a Rashomon set $\mathcal{M}$ (Adebayo et al. 2018; Zhang et al. 2018; Fresz, Lörcher, and Huber 2024). AXE is inspired by methods proposed to evaluate such explanations (Hooker et al. 2018; Fresz, Lörcher, and Huber 2024).

Ground-truth Metrics promote the same Explanation Across Models in a Rashomon Set

Real world situations lack access to an oracle to provide ground-truth explanations $\mathbf{e}^*$ (Zhou et al. 2021; Molnar, Casalicchio, and Bischl 2020). 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 (Krishna et al. 2024; Agarwal et al. 2024). Figure 4 depicts such an example, showcasing a violation of the *model relativism* principle.

Consider a model m with two input features, X_1 and X_2, which makes predictions y, parameterized $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. A related model m' is part of the same Rashomon set $\{m, m'\} \subseteq \mathcal{M}$, has slightly different decision boundaries but makes similar predictions $y \approx y' = \beta'_0 + \beta'_1 X_1 + \beta'_2 X_2$. For datapoint $\mathbf{x}$ and model m feature importances are denoted 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 $\mathbf{e}^* = [\beta_1, \beta_2]$ (Krishna et al. 2024). This comparison takes many forms, with definitions provided in table 1. For 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 $e = [i_1, i_2]$, model m with $\beta_1 = 0.7$ and $\beta_2 = 0.3$ (so $\beta_1 > \beta_2 > 0$), we see that regardless of the specific value



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

Figure 4: **Explanation Quality is Invariant to Model Selection within the Rashomon Set**: 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 $\mathbf{e}^* = (\beta_1, \beta_2) = (0.7, 0.3)$. Consider model $m'(\mathbf{x}) = \beta'_0 + \beta'_1 X_1 + \beta'_2 X_2$ that makes "mostly" similar predictions and belongs to the same Rashomon set $\{m, m'\} \subseteq \mathcal{M}$. If $\beta'_1 > \beta'_2 > 0$, then these explanation quality measurements remain unchanged (despite changes in ground-truth $\mathbf{e}^*$ against which explanation $\mathbf{e}$ will be compared). This violates the *model relativism* principle.

of $\mathbf{e}$, there are regions where the resulting $RA_{n=2}$ is the same (figure 4 a). Similarly, the $SA_{n=2}$ is the same across i_1, i_2 regions (figure 4 b) and $SRA_{n=2}$ too (figure 4 c). Concretely, any explanation $e = (i_1, i_2)$ such that $i_1 > 0$, $i_2 > 0$, and $i_1 > i_2$ (region labeled 1 in figure 4 c) is guaranteed to have the same FA, RA, SA, SRA, RC, and PRA.

Any model $m' \in \mathcal{M}$ with similar predictions such that $m \neq m'$ with the same explanation quality evaluations as model m from figure 4 would violate the *model relativism* principle. Upon observation, all models m' such that $\beta'_1 > \beta'_2 > 0$ satisfy this condition. The relative importance of X_1 with respect to X_2 can range from 1 to ∞ in the limit, but the ground-truth based explanation quality metrics FA, RA, SA, SRA, RC, and PRA would be unchanged. For instance if m' has similar predictions on the test set to m with $\beta'_1 = 0.5$ and $\beta'_2 = 0.3$, then the models belong to the same Rashomon set but the evaluation promotes the exact same explanation for both models, despite the relative importance of the features being different.

This example demonstrates *local contextualization* violations too. The presence of entire regions (as opposed to individual points) in figure 4 with a common explanation quality value q indicates that different possible explanations $\mathbf{e}$ map to the same quality of 1, 0.5, or 0. Since the comparison is always with a fixed vector $\mathbf{e}^*$ which does not change with datapoint $\mathbf{x}$, explanations for different datapoints are incentivized to be identical, promoting a single "global" explanation over per-datapoint "local" explanations.

A Sensitivity and Ground-Truth Agnostic Explanation Evaluation Framework

Inspired by desiderata from user studies (Colin et al. 2022; Alqaraawi et al. 2020; Buçinca et al. 2020; Chandrasekaran et al. 2018; Hase and Bansal 2020; Poursabzi-Sangdeh et al. 2021), and seeking to alleviate common problems with ground-truth based evaluation, 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 (Hooker et al. 2018) 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 (Agarwal et al. 2024).

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 $\mathbf{x}$ (Fix and Hodges 1989; Cover and Hart 1967).

Algorithm 1 defines our framework for evaluating explanations. We denote the target variable Y, predicted from the input dataset $\mathcal{X}$ consisting of ν 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

Algorithm 1: Evaluating Explanation Quality with AXE_n^k

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

```
Initialize an empty list: Ŷ ← []
for each datapoint x_i and explanation e_i in (X, E) do
    Find the n most important input features: f_imp ← ImpFeatures(e_i, n)
    Create X_f with subset of features f_imp from X
    Train K-NN model M_i^k with inputs X_f and target Y_preds
    Obtain prediction ŷ_i from M_i^k for datapoint x_i
    Append ŷ_i to Ŷ
end for
Return performance measure: Accuracy(Ŷ, Y_preds)
```

M_i^k to predict model outputs Y_{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 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. Consequently, 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 further satisfies the *model relativism* principle by training model M_i^k to predict the classifier response Y_{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.

Detecting Explanations Fairwashed Using Different Models from a Rashomon Set

We simulate a state-of-the-art adversarial attack (Slack et al. 2019) 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. The discriminatory decision procedure employed internally by model m remains unchanged in models m_S and m_L, and the three models make similar predictions. The models m, m_S, and m_L therefore form a Rashomon set of models that make similar predictions.

This attack fairwashes model explanations by hiding discriminatory model behavior. Imagine explanations for the diabetes prediction model figure 2 showing that the model

Table 2: **Detecting explanation fairwashing**: We replicate an adversarial fairwashing attack (Slack et al. 2019) and generate spurious explanations, which we try to then detect using quality metrics q that do not need ground-truths: AXE, PGI, and PGU. E_ρ is a set of explanations that correctly denote that the most important model input feature is X_ρ. E_ϕ and E_ψ are explanations where the most important feature is X_ϕ or X_ψ. We are explaining model m_e which uses the same internal mechanism as model m for all datapoint in the data input manifold, ie $\{m, m_e\} \subseteq \mathcal{M}$ and both use feature X_ρ to make predictions. A good evaluation metric q should distinguish manipulated explanations from correct explanations: (i) $\bar{q}(E_\rho) > \bar{q}(E_\phi)$ and (ii) $\bar{q}(E_\rho) > \bar{q}(E_\psi)$; where $\bar{q}(E) = \sum_{\mathbf{e} \in E} q(\mathbf{e})/|E|$. **AXE always scores the correct explanation better than other explanations.**

Dataset	Adversarial Model	Eval. Metric q ($n = 1$)	*Evaluating explanations with a single important attribute:* Protected $\bar{q}(E_\rho)$	Foil 1 $\bar{q}(E_\phi)$	Foil 2 $\bar{q}(E_\psi)$	Other $\bar{q}(E_\omega)$	$\bar{q}(E_\rho) > \bar{q}(E_\phi)$ *and* $\bar{q}(E_\rho) > \bar{q}(E_\psi)$
German Credit	m_{L1} (LIME, 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_{S1} (SHAP, 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_{L1} (LIME, 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_{S1} (SHAP, 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_{L2} (LIME, 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_{S2} (SHAP, 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_{L1} (LIME, 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_{S1} (SHAP, 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_{L2} (LIME, 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_{S2} (SHAP, 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	✔

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_ρ), but they fool explainers SHAP and LIME respectively into generating explanations showing spurious "foil" features (X_ϕ, X_ψ) as the most important (Slack et al. 2019). A good evaluation metric would identify the same explanation to be of high quality for all three models. This explanation would ideally be the "correct" one indicating that a protected feature (X_ρ) was important to the model prediction. Models m_S and m_L are manipulated such that other explanations (showing features $X_\phi or X_\psi$ as important) get generated, but a good explanation evaluation metric q should identify these manipulated explanations by scoring them poorly.

Formally, given a data manifold $\mathcal{D}$ and model m that uses a "protected" input feature X_ρ (eg. race or gender) to make predictions, the attack enumerates models m_e in the Rashomon set of the model $\{m, m_e\} \subseteq \mathcal{M}$ such that they make the same predictions using the same protected input feature on the data manifold, that is $m_e(\mathbf{x}) \approx m(\mathbf{x}) \forall \mathbf{x} \in \mathcal{D}$. The model m_e might differ from m off-manifold, and in the case of m_S and m_L is explicitly designed such that the behavior outside of data manifold $\mathcal{D}$, in the explanation manifold of SHAP or LIME, presents spurious "foil" input features X_ϕ or X_ψ as most important.

Model behavior differs only for datapoints that lie off the data manifold $x \notin \mathcal{D}$. If we were considering these datapoints, where models genuinely used different features (eg. X_ϕ vs X_ψ), then *model relativism* would require explanations to differ: explaining m should indicate protected input feature X_ρ as important, while explaining m_e should ascribe importance to X_ϕ or X_ψ. However, when considering datapoints on-manifold $\mathbf{x} \in \mathcal{D}$, *on-manifold evaluation* requires feature importance explanations for m_e to correctly identify the protected attribute X_ρ as important. Recall that for these datapoints, both m and m_e use the same internal mechanism, making identical predictions using the protected feature X_ρ.

We use the same datasets and models as the original adversarial attack (Slack et al. 2019): the German Credit dataset from lending (Hofmann 1994) and the COMPAS (Larson et al. 2016) dataset and the Communities and Crime dataset (Redmond 2002) from criminal justice. We then manually construct competing feature-importance explanations: E_ρ is the set of all explanations where the protected feature X_ρ is the most important feature. E_ϕ and E_ψ are the explanation sets where the foil feature X_ϕ or X_ψ is the most important, respectively. Finally, E_ω 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_ρ) are correct. Explanations that mask this (E_ϕ, E_ψ) 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 $\bar{q}(E) = \sum_{\mathbf{e} \in E} q(\mathbf{e})/|E|$.

We test AXE against PGI and (-)PGU (PGU values flipped negative so that higher scores are better like PGI and AXE) by scoring various explanations for models m_e. We use the German Credit (Hofmann 1994) dataset from lending and the COMPAS (Larson et al. 2016) and Communities and Crime (Redmond 2002) datasets from criminal justice. We assume the adversary is interested in deploying binary classifier m that makes predictions using protected characteristics X_ρ race or gender. We consider 4 kinds models in the Rashomon set: designed to fool explainers SHAP and LIME, and with 1 or 2 spurious foil features respectively, giving us $\mathcal{M} = \{m_{S1}, m_{S2}, m_{L1}, m_{L2}\}$. We then score competing explanations for each of these models, checking that explanations identifying the protected feature X_ρ as most important receive the highest scores.

Table 2 we summarizes our results. To test explanations constructed to consider one feature as important and all others as unimportant, we use AXE $_{n=1}$, PGI$_{n=1}$ and PGU$_{n=1}$. The last column 2 shows PGU failing to discern correct explanations E_ρ from spurious ones E_ψ, E_ϕ 7 out of 10 times, and PGI failing 3 out of 10 times. The overall error rate for PGI and PGU is 50%, while AXE has no failures yielding an error-rate of 0%. Off-manifold manipulation by models m_e from the Rashomon set of the model $m \in \mathcal{M}$, designed to mislead explainers SHAP and LIME, also misleads sensitivity based evaluation metrics PGI and PGU.

Conclusion

In this paper we have focused on two important aspects of XAI as they relate to the Rashomon effect. If models make similar predictions but through different internal mechanisms, explanations should seek to highlight these differences (*model relativism*). Figure 4 presents an example case where prior ground-truth based metrics fail this simple test, and promote a single explanation across models in a Rashomon set. Conversely, if models in a Rashomon set use the same internal mechanism to make a prediction, differing only off-manifold, then explanations on-manifold should be the same for all models (*on-manifold evaluation*). For models from a Rashomon set that have the exact same internal mechanisms, if explanations generated are different due to adversarial fairwashing then this should be detected when evaluating them. Our experiment in table 2 shows that unlike prior metrics, our proposed metric AXE can detect adversarial fairwashing because it follows the three principles of explanation evaluation. While this abridged paper highlights connections between explanation evaluation and predictive multiplicity caused by the Rashomon effect, additional experiments and theory situating AXE in the AI explainability literature can be found in the original paper "Evaluating Model Explanations without Ground-truth".

Acknowledgments

This research was supported 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. This work has also been supported by the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship (Humboldt Professor of Technology and Regulation) endowed by the Federal Ministry of Education and Research via the Hasso Plattner Institute. The donors had no role in the decision to publish or the preparation of this paper.

References

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

Agarwal, C.; Krishna, S.; Saxena, E.; Pawelczyk, M.; Johnson, N.; Puri, I.; Zitnik, M.; and Lakkaraju, H. 2024. OpenXAI: towards a transparent evaluation of post hoc model explanations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Alqaraawi, A.; Schuessler, M.; Weiß, P.; Costanza, E.; and Berthouze, N. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, 275–285. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371186.

Barr, B.; Fatsi, N.; Hancox-Li, L.; Richter, P.; Proano, D.; and Mok, C. 2023. The Disagreement Problem in Faithfulness Metrics. arXiv:2311.07763.

Buçinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, 454–464. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371186.

Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8).

Chandrasekaran, A.; Prabhu, V.; Yadav, D.; Chattopadhyay, P.; and Parikh, D. 2018. Do explanations make VQA models more predictable to a human? In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1036–1042. Brussels, Belgium: Association for Computational Linguistics.

Chang, V.; Bailey, J.; Xu, Q. A.; and Sun, Z. 2022. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput Appl*, ( ): 1–17.

Colin, J.; FEL, T.; Cadene, R.; and Serre, T. 2022. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 2832–2845. : Curran Associates, Inc.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27.

Dai, J.; Upadhyay, S.; Aivodji, U.; Bach, S. H.; and Lakkaraju, H. 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*, AIES '22, 203–214. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Fel, T.; Cadène, R.; Chalvidal, M.; Cord, M.; Vigouroux, D.; and Serre, T. 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. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713845393.

Fix, E.; and Hodges, J. L. 1989. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3): 238–247.

Fong, R.; Patrick, M.; and Vedaldi, A. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2950–2958. : .

Fong, R. C.; and Vedaldi, A. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 3449–3457. : .

Fresz, B.; Lörcher, L.; and Huber, M. 2024. Classification Metrics for Image Explanations: Towards Building Reliable XAI-Evaluations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1–19. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 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. : .

Hase, P.; and Bansal, M. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5540–5552. Online: Association for Computational Linguistics.

Hofmann, H. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Hooker, S.; Erhan, D.; jan Kindermans, P.; and Kim, B. 2018. Evaluating Feature Importance Estimates. *arXiv*, ( ): .

Ivanovs, M.; Kadikis, R.; and Ozols, K. 2021. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recogn. Lett.*, 150(C): 228–234.

Kapishnikov, A.; Bolukbasi, T.; Viegas, F.; and Terry, M. 2019. XRAI: Better Attributions Through Regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4947–4956. : .

Krishna, S.; Han, T.; Gu, A.; Wu, S.; Jabbari, S.; and Lakkaraju, H. 2024. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *Transactions on Machine Learning Research*, ( ): .

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How We Analyzed the COMPAS Recidivism Algorithm.

Liao, Q. V.; and Varshney, K. R. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *CoRR*, abs/2110.10790.

Liu, Y.; Khandagale, S.; Khandagale, S.; White, C.; and Neiswanger, W. 2021. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. : .

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive Multiplicity in Classification. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6765–6774. : PMLR.

Mhasawade, V.; Rahman, S.; Haskell-Craig, Z.; and Chunara, R. 2024. Understanding Disparities in Post Hoc Machine Learning Explanation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2374–2388. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Molnar, C.; Casalicchio, G.; and Bischl, B. 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In Koprinska, I.; Kamp, M.; Appice, A.; Loglisci, C.; Antonie, L.; Zimmermann, A.; Guidotti, R.; Özgöbek, Ö.; Ribeiro, R. P.; Gavaldà, R.; Gama, J.; Adilova, L.; Krishnamurthy, Y.; Ferreira, P. M.; Malerba, D.; Medeiros, I.; Ceci, M.; Manco, G.; Masciari, E.; Ras, Z. W.; Christen, P.; Ntoutsi, E.; Schubert, E.; Zimek, A.; Monreale, A.; Biecek, P.; Rinzivillo, S.; Kille, B.; Lommatzsch, A.; and Gulla, J. A., eds., *ECML PKDD 2020 Workshops*, 417–431. Cham: Springer International Publishing. ISBN 978-3-030-65965-3.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s).

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020*, WWW '20, 3126–3132. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370233.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. *CoRR*, abs/1806.07421: .

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Wortman Vaughan, J. W.; and Wallach, H. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Qiu, L.; Yang, Y.; Cao, C. C.; Zheng, Y.; Ngai, H.; Hsiao, J.; and Chen, L. 2022. Generating Perturbation-based Explanations with Robustness to Out-of-Distribution Data. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 3594–3605. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Redmond, M. 2002. Communities and Crime. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53W3X.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

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

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. : .

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2019. How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods. *CoRR*, abs/1911.02508: .

Zhang, J.; Bargal, S. A.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2018. Top-Down Neural Attention by Excitation Backprop. *Int. J. Comput. Vision*, 126(10): 1084–1102.

Zhou, J.; Gandomi, A. H.; Chen, F.; and Holzinger, A. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5): .