Resolving Disagreement Problems in Explainable Artificial Intelligence Through Multi-Criteria Decision Analysis

Anonymous authors
Paper under double-blind review

Abstract

Post-hoc explanation methods are critical for building trust in complex black-box artificial intelligence (AI) models; however, they often suffer from the disagreement problem, which provides conflicting explanations for the same prediction. This inconsistency undermines reliability and poses a significant barrier to adoption in high-stakes domains that demand trustworthiness and transparency. To address this, we move beyond the search for a single best method and instead propose a principled, preference-driven framework for selecting the best suitable explanation technique for a given context: *which specific post-hoc explanation methods to use and when?* We formalize this selection process as a Multi-Criteria Decision Analysis (MCDA) problem. Our framework evaluates a set of state-of-the-art post-hoc explanation methods (e.g., LIME, SHAP, and Anchor) against six explanation evaluation metrics: fidelity, identity, stability, separability, faithfulness, and consistency. We then apply a suite of established MCDA techniques such as Simple Additive Weighting (SAW), Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), and Elimination and Choice Translating Reality (ELECTRE I) to aggregate these evaluations based on user-defined priorities. By comparing the rankings produced by these diverse decision logics across multiple predictive models and real-world datasets, we demonstrate not only how to select the optimal explanation method under different priority scenarios (e.g., favoring fidelity vs. stability) but also how to expose critical trade-offs that are invisible to simpler aggregation approaches. Our work provides a robust, transparent, and adaptable methodology for resolving explanation disagreement, enabling practitioners to make more justifiable choices about explainability.

1 Introduction

The deployment of complex machine learning (ML) and deep learning (DL) models in high-stakes domains, such as finance, law, and healthcare, hinges on our ability to trust their decisions Davenport & Kalakota (2019); Cao (2020). Post-hoc explainable AI (XAI) methods, such as LIME Ribeiro et al. (2016), SHAP Lundberg & Lee (2017), etc., are crucial tools for building trust by offering insights into the rationale behind ML/DL model predictions. However, these methods introduce a significant paradox: while designed to enhance the transparency of ML/DL models, they often generate confusion. When applied to the same prediction, different XAI methods frequently highlight different features as most important, leading to what is known as the *disagreement problem* Krishna et al. (2022); Han et al. (2022); Silva et al. (2025); Laberge et al. (2024). Let us consider a healthcare scenario in which a predictive ML/DL model is being deployed to aid doctors in diagnosing diabetic diseases by analyzing a range of patient data (e.g., medical history, genetic information, lifestyle factors, various clinical measurements, etc.,), as shown in Figure 1. In addition, to enhance the transparency and trustworthiness of this ML/DL-driven diagnostic process, state-of-the-art (SOTA) post-hoc explanation methods (e.g., SHAP and LIME) are employed to explain each model prediction to a doctor. These methods generate detailed explanations to elucidate the reasoning behind the ML/DL model's predictions, helping doctors make informed decisions. However, an issue arises when different explanation methods are applied to the same diagnostic task, each focusing on a different set of features to explain the ML/DL model's predictions. For instance, SHAP identifies features such as blood sugar levels, physical activity, etc., that cause diabetes. At the same time, LIME highlights features such as pregnancies, smoking history, etc., that cause diabetes. Such disparities can lead to uncertainty among doctors and undermine the trustworthiness of the AI-based diagnostic system. The presence of conflicting explanations can lead doctors to doubt the model's recommendations, as they are unsure which explanation to prioritize or trust. This disagreement may lead doctors to rely more on their own clinical expertise, potentially resulting in inconsistent or suboptimal decisions in some instances. Thus, to ensure the trustworthiness of the overall decision-making process, it becomes essential to address these discrepancies among explanation methods. Standardizing or combining these explanations into a coherent, unified narrative could mitigate trust issues and enhance collaborative decision-making among AI systems and professionals across sectors (e.g., healthcare, finance), raising a critical unanswered question: *which explanation method should I trust?*

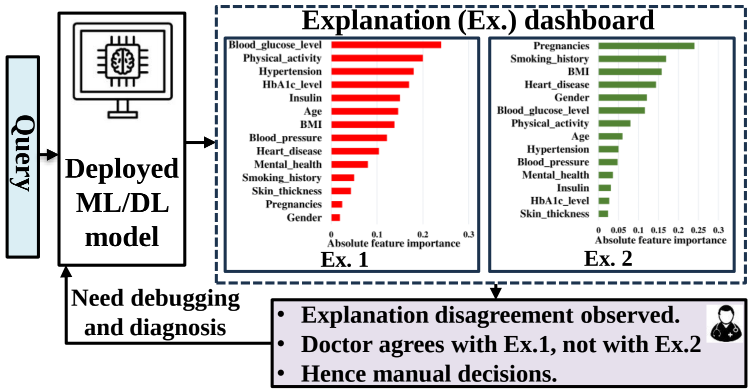


Figure 1: An illustrative example highlights how disagreement in explanation leads to uncertainty among doctors and impacts the trustworthiness of the AI-based diagnostic system.

The root of this disagreement lies in the diverse theoretical foundations of these XAI methods, ranging from game-theoretic Shapley values to local surrogate modeling. To evaluate the quality of these methods, researchers have proposed numerous explanation evaluation metrics assessing desirable properties, such as *fidelity* (how well the explanation reflects the model), *stability* (robustness to minor perturbations) *faithfulness* (if features truly drive the prediction), etc., Alvarez-Melis & Jaakkola (2018); Zhou et al. (2021); Agarwal et al. (2022); Klein et al. (2024). However, this has led to another dilemma: no single XAI method consistently performs well across all metrics. One method may demonstrate high fidelity but poor stability, while another offers greater robustness at the expense of faithfulness. As a result, practitioners face a multi-criteria dilemma, lacking a principled framework to navigate these trade-offs and select the most *suitable explanation method* for their specific context and objectives. The critical research gap, therefore, is not the absence of evaluation metrics but rather the lack of a structured decision-making process to effectively synthesize them.

To address this gap, we formalize the selection of an appropriate explanation method as a *Multi-Criteria Decision Analysis (MCDA)* problem. We propose a novel framework that moves beyond the futile search for a single, universally *"best"* explainer, and instead provides a transparent, preference-driven process for selecting the most appropriate one. Instead of relying on a single aggregation strategy, our core contribution lies in a comparative analysis of three established MCDA techniques with distinct decision-making philosophies: the compensatory Simple Additive Weighting (SAW), the ideal-point-based Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), and the non-compensatory outranking method ELECTRE I. By applying these techniques, our framework recommends an explanation method that best aligns with user-defined priorities (e.g., favoring fidelity over stability) and, importantly, exposes the inherent trade-offs and sensitivity of the decision. Our key contributions are threefold:

- At first, we empirically show that SOTA post-hoc explanation methods (e.g., LIME, SHAP, Anchor) exhibit significant performance trade-offs across six key explanation evaluation metrics: fidelity, identity, stability, separability, faithfulness, and consistency, reinforcing the need for a structured decision framework.

- We then propose a novel and adaptable framework that formulates the explanation method selection task as an MCDA problem, enabling user-defined preferences (weights) to guide the decision-making process. Specifically, we apply three MCDA techniques: SAW, TOPSIS, and ELECTRE I to diverse models and datasets, demonstrating how our framework not only provides a suitable explanation method but also uncovers deeper insights into the decision logic, such as when compensatory vs. non-compensatory reasoning leads to different outcomes.

Evaluated across four ML/DL models, three explanation methods, and three real-world datasets, our approach provides a robust and principled framework for resolving XAI disagreements in a transparent and justifiable manner. For instance, across all datasets, SHAP consistently ranks higher in most scenarios for the predicted class explanations. These findings underscore the often-overlooked issue of explanation disagreement and offer practical guidance for selecting appropriate explainers for predictive models.

2 Related Works

In this section, we provide an overview of the current evaluation paradigm in XAI research. Thus, our work builds on the vast literature in XAI. We review prior work and its connections to this research.

SOTA XAI methods have two main approaches: (i) designing inherently interpretable models such as rule lists, decision trees, etc., (ii) applying post-hoc explanation methods, i.e., SHAP, LIME, etc., to explain a black-box ML/DL model locally (for a specific sample) or globally (for the entire model space). Our work focuses on local post-hoc explanation methods based on feature importance. For instance, post-hoc explainers, such as LIME Ribeiro et al. (2016), SHAP Lundberg & Lee (2017), Anchor Ribeiro et al. (2018), BayesLIME Slack et al. (2021), BayesSHAP Slack et al. (2021), etc., are proxy models trained atop a base ML/DL model with the sole intention of *"explaining"* that base model. These methods rely solely on the model's inputs and outputs to identify the salient features that contribute to its output. On the other hand, methods such as SmoothGrad Smilkov et al. (2017), Integrated Gradients Qi et al. (2019), GradCAM Selvaraju et al. (2017), etc., are referred to as gradient-based local explanation methods, which do not use a proxy model but instead compute the gradients of a model with respect to input features to identify important features. A more detailed treatment of this topic is provided in other comprehensive survey articles Madsen et al. (2022); Zhao et al. (2023). However, one problem with these explanation methods is that they often suffer from the *disagreement problem* Krishna et al. (2022); Fazelpour & Fleisher (2025); Silva et al. (2025). From a conceptual standpoint, the misalignment of goals among explanation methods leads to an inconsistent view of explanations Reingold et al. (2024). For instance, the SHAP method is based on game theoretic concepts Lundberg & Lee (2017), whereas LIME is motivated by the function approximation method Ribeiro et al. (2016). Such differences pose conceptual and practical challenges for understanding and using explanation methods, thwarting progress in the XAI field and raising questions about which explanation method to use and when.

Prior research has proposed several metrics to assess the reliability of an explanation. An extensive survey of metrics for evaluating explanation methods is provided in Zhou et al. (2021); Klein et al. (2024); Agarwal et al. (2022). Their survey highlights two high-level goals of explanation methods: interpretability (the clarity, simplicity, and broadness of the explanations) and fidelity (the completeness and soundness of the explanations). In contrast, authors in Liu et al. (2021) proposed a synthetic benchmark for explanation evaluation, including implementations of several metrics and a discussion of how to choose evaluation metrics. Neely et al. Neely et al. (2021) measured the disagreement problem in XAI for the first time, in which they compared different explanation methods (e.g., LIME, DeepSHAP, etc.,) with a rank correlation (Kendall's τ) metric. Instead of only using a rank correlation metric, authors in Krishna et al. (2022) used different evaluation metrics (e.g., feature agreement, rank agreement, sign agreement, etc.,) to capture specific aspects of explanation disagreement w.r.t. the top-k features using various types of data. Unfortunately, their proposed approach did not address how to choose an accurate and trustworthy explanation for the predictive model's decisions. This is important since unstable and inconsistent explanations may lead to an untrustworthy model for end users (e.g., doctors and policymakers). Therefore, Han et.al Han et al. (2022) extend the work of Krishna et al. Krishna et al. (2022) to investigate why the disagreement problem exists for these explanation methods. They conclude that different explanation methods approximate predictive black-box models over different neighborhoods by applying different loss functions. However, one problem with their work is that it focuses on faithfulness rather than on the interpretability of explanations. On the other hand, Avi et.al. Schwarzschild et al. (2023) showed the disagreement problem using the Post hoc Explainer Agreement Regularization (PEAR) loss term that measures the difference in feature attribution between a pair of explainers. A more focused disagreement problem study can be found in Roy et al. (2022) where the explanations of LIME and SHAP are investigated for one single defect prediction model. They conclude that LIME and SHAP disagree more on the ranking of important features than on their signs of importance. In addition to these evaluation metrics, human-grounded evaluation approaches emphasize how human users perceive and utilize explanations Linardatos et al. (2020); Schmitt et al. (2024); Doshi-Velez & Kim (2017); Swamy et al. (2025). For instance, authors in Kaur et al. (2020) found that explanations are often over-trusted and misused. Although our work does not involve human subject studies, we view this line of human-centered investigation as a critical complement to our quantitative evaluation framework. Unfortunately, none of these studies considered how to select an accurate explanation method for the black-box ML/DL model's decisions. This is important because unstable, inconsistent explanations may lead to untrustworthy decisions by end users. Furthermore, the quality of the explanations and the corresponding evaluation metrics were used to analyze the behavior of the explanation methods. Since SOTA explanation methods may result in explanations that are not only inconsistent and unstable but also prone to adversarial attacks and potentially mislead end users into believing that the underlying models and explanations are fair Ghorbani et al. (2019); Slack et al. (2020); Madsen et al. (2022). This, in turn, could lead to the deployment of unfair models in critical real-world applications and ought not to be used in high-stakes decisions. Indeed, it is important to understand and quantify how often explanation methods disagree, and to examine how ML practitioners and domain experts can address such disparities. This is concerning, as inconsistent or misleading explanations can erode trust and propagate unfair decisions in high-stakes settings Madsen et al. (2022).

3 Methodology

This section presents the details of our proposed methodology for a trustworthy explanation selection from a set of explanation methods. As illustrated in Figure 2, the proposed method is comprised of three phases: (1) Classification model and explanation method development, (2) Explanation evaluation metrics development, and (3) MCDA method. The details are as follows.

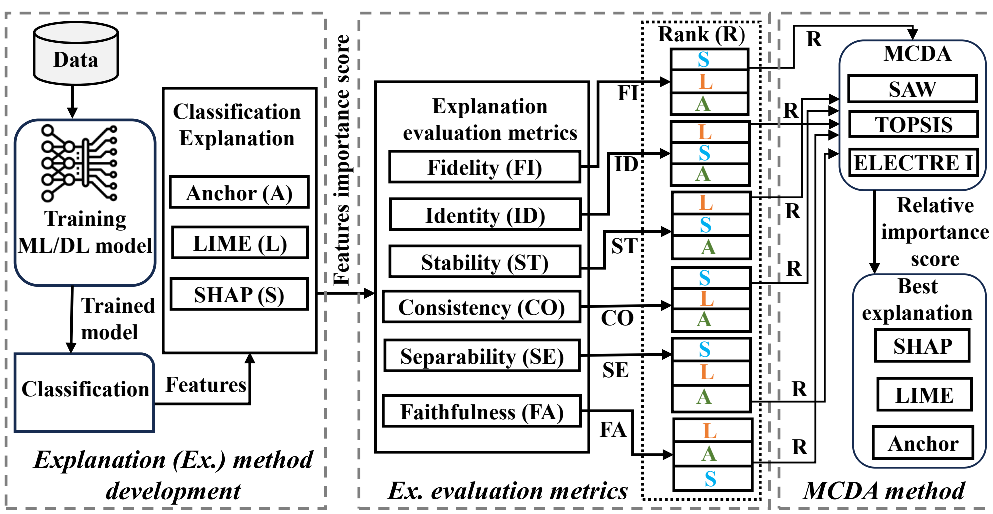


Figure 2: Our proposed framework for selecting an appropriate explanation method from a set of explanation methods.

3.1 Explanation Method Development

At first, we develop a classification model using three ML models, namely Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), and a 3-layer neural network (NN) with 50, 100, and 200 nodes per layer, ReLU activation, binary cross-entropy loss, Adam optimizer (learning rate 0.001), and early stopping strategy (patience value = 30) over 100 epochs. These models are trained on three real-world datasets: HELOC Demajo et al. (2020), German Credit Asuncion & Newman (2007), and COMPAS Angwin et al. (2016). Hyperparameter tuning is performed using a grid search, and performance evaluation is conducted through 10-fold cross-validation. Then, to interpret ML/DL models' predictions, we apply three local post-hoc explanation methods, such as LIME Ribeiro et al. (2016), (SHAP) Lundberg & Lee (2017), and Anchor Ribeiro et al. (2018)to compute feature importance scores for randomly selected test samples. It is important to note that this study focuses solely on local explanation methods, where samples are randomly selected from the test dataset. We adopt widely used ML and DL models, as well as three post-hoc explanation techniques and three datasets, aligning with prior works in the XAI disagreement literature Laberge et al. (2024); Han et al. (2022); Krishna et al. (2022).

3.2 Explanation Evaluation Metrics

Several metrics have been proposed to evaluate the explainability of XAI methods Alvarez-Melis & Jaakkola (2018); Agarwal et al. (2022). A comprehensive survey by Zhou et al. (2021) identifies two key qualities of high-quality explanations: how well it approximates the model and how human-understandable it is. In this work, we focus on six explanation evaluation metrics: fidelity, identity, stability, separability, consistency, and faithfulness, which help measure the explanation disagreement problem in a principled manner. These metrics are commonly used in recent XAI studies to assess explanation reliability across different contexts Barr et al. (2023); Agarwal et al. (2022); Dai et al. (2022); Bobek et al. (2021); Parimbelli et al. (2023); Solís-Martín et al. (2023); Klein et al. (2024). The details of these explanation evaluation metrics are below.

Fidelity: A key question in evaluating explanation quality is *"To what extent does explanation accurately represent the underlying decision-making process?"* Dai et al. (2022). Explanations that correctly identify important features are said to have high fidelity. This is quantified as:

$$F_{(x,g,\epsilon)} = \frac{|top(k, \boldsymbol{V}) \cap top(k, v)|}{k}$$

Here, v denotes the ground truth feature importance weights of a black-box model g for input x, and $\boldsymbol{V} = \epsilon(x, g)$ is the explanation generated by explanation method ϵ, the fidelity metric F measures the fraction of top-k features in $\boldsymbol{V}$ that overlap with those in v, where k selects the indices of the k largest elements. The intuition behind this fidelity metric is that the model's prediction, g, should remain stable when non-important features (outside the top-k) are perturbed. Significant changes indicate low fidelity, revealing that the explanation fails to capture the essential features critical to the model's prediction.

Stability: This metric, also known as explanation robustness, captures the idea that similar inputs should yield similar explanations. We use the Lipschitz indicator proposed by Alvarez-Melis and Jaakkola Alvarez-Melis & Jaakkola (2018) to quantify stability, denoted as:

$$S_X(x_i) = \max_{x_j \in N_\varepsilon(x_i)} \frac{||M_i - M_j||}{||x_i - x_j||}$$

Here, $S_X(x_i)$ denotes the stability of test sample $x_i \in X$, with M_i and M_j representing feature importance scores for x_i and its neighbor x_j, respectively. The neighborhood $N_\varepsilon(x_i)$ includes points within L2 distance ϵ of x_i. Intuitively, stability measures how much explanations change under small input perturbations; lower Lipschitz values indicate higher robustness at point x_i Alvarez-Melis & Jaakkola (2018)

Faithfulness: An explanation is faithful if the features it highlights genuinely influence the model's prediction. One way to measure this is through the Prediction Gap on Important features (PGI), which quantifies the change in model output when top-k important features (as identified by an explanation) are perturbed Agarwal et al. (2022). A larger gap implies higher faithfulness. PGI is defined as: $U_{(x,g)} = \frac{1}{p} \sum_{i=1}^{p} |g(x) - g(\tilde{x}_i)|$, where $U_{(x,g)}$ is the faithfulness score for input x, $\tilde{x}_i$ is the perturbed version of x (top-k features modified), g is the black-box model, and p is the number of stochastic perturbation runs. Following Barr et al. (2023), we apply Gaussian noise ($\mathcal{N}(0, 0.1)$) for continuous features and flip 30% of discrete ones. Note, this work uses only the PGI metric to evaluate faithfulness Agarwal et al. (2022).

Consistency: This metric assesses how similar the explanations from different methods are for the same prediction Bobek et al. (2021). The main intuition is that repeated explanations for a single observation should produce similar results; otherwise, it suggests instability in the ML/DL model or the explanation method. Consistency is computed by averaging the L1 distance between pairs of explanations, discussed in Bobek et al. (2021). For a given observation, the consistency score is defined as:

$$C(M_{j_1}, M_{j_2}) = \frac{1}{||M_{j_1} - M_{j_2}||_2 + 1}$$

Where $C(M_{j_1}, M_{j_2})$ measures the consistency for the j^{th} observation, and M_{j_1}, M_{j_2} the feature importance from two explanation methods.

Identity: The underlying assumption is that if there are two identical data points(i.e., the actual and predicted classes) their explanations must be identical Parimbelli et al. (2023). If this condition is not met, then the black-box ML/DL model either predicted the incorrect class or the explanation method produced an explanation that is not identical. The identity metric can be expressed as follows: $\forall_{i,j}(d(x_i, x_j) = 0 \Rightarrow d(\epsilon_i, \epsilon_j) > 0)$. Where x are samples, d is a distance function, and ϵ is the explanation vector that explains the prediction of each sample.

Separability: This metric evaluates whether dissimilar inputs produce dissimilar explanations Parimbelli et al. (2023); Solís-Martín et al. (2023). The underlying assumption is that non-identical samples should not receive identical explanations, as each feature is expected to contribute meaningfully either positively or negatively to the model's prediction. For instance, if two samples differ only in a feature that does not influence the prediction, the model may produce the same output, and the explanation method might produce identical explanations. Formally, this can be expressed as: $\forall_{i,j}(d(x_i, x_j) \neq 0 \Rightarrow d(\epsilon_i, \epsilon_j) = 0)$, where x are input samples, d is a distance function, and ϵ is the explanation vector for each sample.

3.3 Proposed Comparative MCDA Framework for XAI Method Selection

To address the challenge of selecting an appropriate explanation method when faced with disagreements, we propose a novel framework based on MCDA. This framework allows for a structured and transparent evaluation of explanation methods against multiple explanation evaluation metrics, with the flexibility to incorporate user-defined preferences through differential weighting. Instead of relying on a single aggregation rule, we implement and compare three established MCDA techniques: Simple Additive Weighting (SAW), Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), and ELECTRE I.

3.3.1 Justification of Methodological Choices

Our framework's components are chosen to ensure a holistic and robust analysis. We selected six evaluation criteria to cover three key dimensions of explanation quality: **Model Alignment** (Fidelity, Faithfulness), **Robustness** (Stability, Identity, Separability), and **Inter-Method Agreement** (Consistency). For the decision logic, we chose three distinct and widely recognized MCDA families to analyze the sensitivity of the rankings to the aggregation philosophy: **SAW** as a simple compensatory baseline, **TOPSIS** as an ideal-point method, and **ELECTRE I** as a non-compensatory outranking method capable of identifying incomparable alternatives and veto effects often missed by other techniques.

3.3.2 Formal Preliminaries

Let the set of candidate explanation methods be $A = \{A_1, \ldots, A_m\}$, where m is the total number of methods. Let the set of evaluation criteria be $C = \{C_1, \ldots, C_n\}$, where n is the total number of metrics. The first step of our framework, detailed in Algorithm 1, is to construct a normalized decision matrix $\mathbf{X}$. The performance of each method A_j on criterion C_i is captured by the element $x_{ij} \in [0, 1]$ in this matrix, where a higher value is always preferable. This is achieved by first computing a raw performance matrix $\mathbf{X}' = [x'_{ij}]$ and then applying two transformations. Any "cost" criterion (where lower is better) is converted to a benefit criterion using Eq. 1, and all scores are then normalized using min-max scaling as shown in Eq. 2:

$$x''_{ij} = \frac{\max_k(x'_{ik}) - x'_{ij}}{\max_k(x'_{ik}) - \min_k(x'_{ik})} \tag{1}$$

$$x_{ij} = \frac{x''_{ij} - \min_k(x''_{ik})}{\max_k(x''_{ik}) - \min_k(x''_{ik})} \tag{2}$$

The user's preferences are captured in a weight vector $W = \{w_1, \ldots, w_n\}$, where w_i is the non-negative weight for criterion C_i and $\sum w_i = 1$.

3.3.3 Formal Properties of the Framework

A robust selection framework should satisfy desirable formal properties. We define Monotonicity and show that our compensatory methods adhere to it.

Definition 1 (Monotonicity). A selection framework is monotonic if improving the performance of an alternative A_j on any single criterion C_i, while holding all other scores constant, does not worsen its final rank. Formally, if $S(A_j)$ is

Algorithm 2 Compensatory MCDA Ranking (SAW & TOPSIS)

Input: Decision Matrix $\mathbf{X}$, Weight Vector W.
Output: SAW Ranking R^{SAW}, TOPSIS Ranking R^{TOPSIS}.

// SAW Calculation
for all $A_j \in A$ **do** calculate $S_j^{\text{SAW}} = \sum_{i=1}^{n} w_i x_{ij}$.
  $R^{\text{SAW}} \leftarrow$ rank alternatives by S_j^{SAW} (descending).
  // TOPSIS Calculation
  Construct weighted matrix $\mathbf{V} = [w_i x_{ij}]$.
  Determine PIS A^+ and NIS A^- from $\mathbf{V}$.
  for all $A_j \in A$ **do** calculate distances D_j^+ to A^+ and D_j^- to A^-.
    for all $A_j \in A$ **do** calculate $CC_j^{\text{TOPSIS}} = D_j^- / (D_j^+ + D_j^-)$.
      $R^{\text{TOPSIS}} \leftarrow$ rank alternatives by CC_j^{TOPSIS} (descending).
      return $R^{\text{SAW}}, R^{\text{TOPSIS}}$.

the final score for alternative A_j, then for any two decision matrices $\mathbf{X}$ and $\hat{\mathbf{X}}$ that are identical except that $\hat{x}_{ij} > x_{ij}$ for some i, j, it must hold that $S(A_j)_{\hat{\mathbf{X}}} \geq S(A_j)_{\mathbf{X}}$.

Proof Sketch for SAW. The SAW method satisfies Monotonicity. The score for an alternative A_j is $S_j^{\text{SAW}} = \sum_{k=1}^{n} w_k x_{kj}$. If we increase a single score x_{ij} to $\hat{x}_{ij}$, the new score becomes $\hat{S}_j^{\text{SAW}} = (\sum_{k \neq i} w_k x_{kj}) + w_i \hat{x}_{ij}$. Since $w_i \geq 0$ and $\hat{x}_{ij} > x_{ij}$, it follows that $w_i \hat{x}_{ij} \geq w_i x_{ij}$, and therefore $\hat{S}_j^{\text{SAW}} \geq S_j^{\text{SAW}}$. This ensures that an improvement in any criterion never penalizes an alternative.

Algorithm 1 Data Preparation and Normalization

Input: Datasets (D_s), Models (M_k), XAI Methods (A), Criteria (C).
Output: Normalized Decision Matrix $\mathbf{X}$.

Compute raw performance matrix $\mathbf{X}' = [x'_{ij}]$ for each A_j on C_i.
Initialize transformed matrix $\mathbf{X}''$.
for all criterion $C_i \in C$ **do**
  if C_i is a cost criterion **then**
    Apply cost-to-benefit transformation (Eq. 1) to get column i of $\mathbf{X}''$.
  else
    Copy column i from $\mathbf{X}'$ to $\mathbf{X}''$.
Apply min-max normalization (Eq. 2) to $\mathbf{X}''$ to get $\mathbf{X}$.
return $\mathbf{X}$. =0

Algorithm 3 Non-Compensatory MCDA (ELECTRE I)

Input: Decision Matrix $\mathbf{X}$, Weight Vector W, Thresholds $\bar{c}, \bar{d}$.
Output: Outranking Kernel P^{ELECTRE}.

Initialize outranking graph G.
for all ordered pairs (A_k, A_l) where $k \neq l$ **do**
  Calculate concordance $c(A_k, A_l) = \sum_{i: x_{ik} \geq x_{il}} w_i$.
  Calculate discordance $d(A_k, A_l) = \frac{1}{L} \max_{i: x_{il} > x_{ik}} (x_{il} - x_{ik})$.
  if $c(A_k, A_l) \geq \bar{c}$ AND $d(A_k, A_l) \leq \bar{d}$ **then**
    Add directed edge $A_k \to A_l$ to G.
$P^{\text{ELECTRE}} \leftarrow$ find nodes in G with in-degree of 0 (the kernel).
return P^{ELECTRE}. =0

3.3.4 Simple Additive Weighting (SAW)

The SAW method, also known as the Weighted Sum Model (WSM), is a compensatory technique that calculates a total score for each explanation method. The process is detailed in Algorithm 2. The score, denoted as S_j^{SAW}, for method A_j is calculated as:

$$S_j^{\text{SAW}} = \sum_{i=1}^{n} w_i x_{ij} \tag{3}$$

where w_i is the weight of criterion C_i and x_{ij} is the normalized performance score of method A_j on that criterion. The explanation methods are then ranked based on their S_j^{SAW} scores in descending order.

3.3.5 Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

TOPSIS is an ideal-point method that ranks alternatives based on their simultaneous shortest distance from a best-case solution and farthest distance from a worst-case solution, as detailed in Algorithm 2. The key steps are:

1. Construct the Weighted Normalized Decision Matrix: Using the normalized decision matrix of scores x_{ij} and weights w_i, we construct a weighted matrix where each element v_{ij} is calculated as $v_{ij} = w_i x_{ij}$.

2. Determine Ideal Solutions: The Positive Ideal Solution (PIS), denoted A^+, is a vector $\{v_1^+, \ldots, v_n^+\}$ where each element v_i^+ is the maximum score achieved on criterion C_i across all explanation methods (i.e., $v_i^+ = \max_j v_{ij}$). Conversely, the Negative Ideal Solution (NIS), denoted A^-, is a vector $\{v_1^-, \ldots, v_n^-\}$ where each element v_i^- is the minimum score ($v_i^- = \min_j v_{ij}$).

3. Calculate Separation Measures: For each explanation method A_j, its Euclidean distance from the PIS is calculated as D_j^+, and its distance from the NIS is calculated as D_j^-:

$$D_j^+ = \sqrt{\sum_{i=1}^{n} (v_{ij} - v_i^+)^2} \tag{4}$$

$$D_j^- = \sqrt{\sum_{i=1}^{n} (v_{ij} - v_i^-)^2} \tag{5}$$

4. Calculate Relative Closeness: The final score for each explanation method A_j, known as the closeness coefficient CC_j^{TOPSIS}, is calculated as:

$$CC_j^{\text{TOPSIS}} = \frac{D_j^-}{D_j^+ + D_j^-} \tag{6}$$

where $0 \leq CC_j^{\text{TOPSIS}} \leq 1$. Explanation methods are then ranked in descending order of their CC_j^{TOPSIS} score.

3.3.6 ELECTRE I (Elimination and Choice Translating Reality)

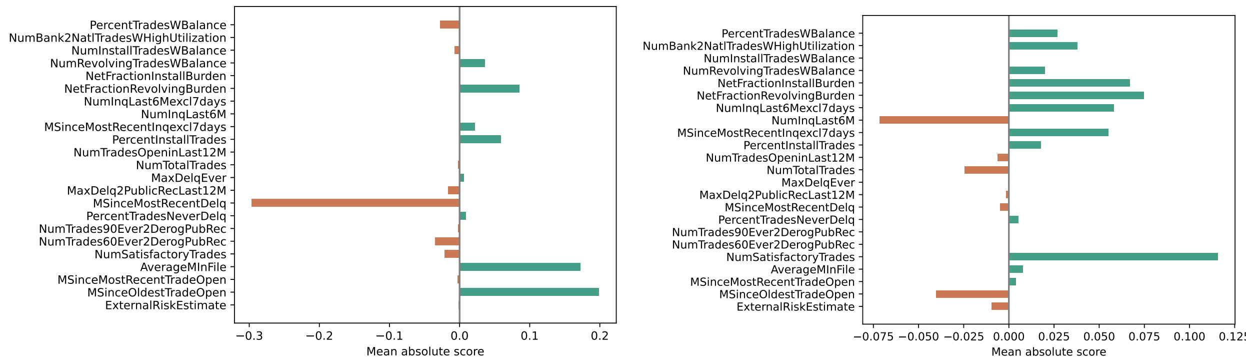
ELECTRE I is a non-compensatory outranking method that builds a preference relation between pairs of alternatives, as detailed in Algorithm 3. It determines if an alternative A_k outranks another alternative A_l (denoted $A_k \succeq A_l$) by constructing concordance and discordance indices.

1. Define Concordance Index: The concordance index, $c(A_k, A_l)$, measures the strength of evidence supporting the claim that A_k is at least as good as A_l. It is the sum of weights for all criteria where A_k performs greater than or equal to A_l:

$$c(A_k, A_l) = \sum_{i: x_{ik} \geq x_{il}} w_i \tag{7}$$

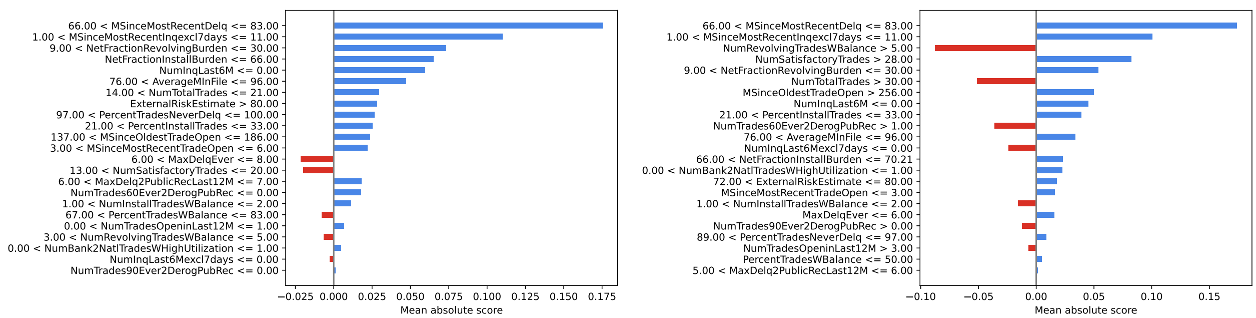
where the summation is over the set of indices i for which the score of explanation method A_k (x_{ik}) is greater than or equal to the score of method A_l (x_{il}).

2. Define Discordance Index: The discordance index, $d(A_k, A_l)$, measures the strength of evidence against the claim that A_k outranks A_l. It identifies the criterion where A_l most significantly outperforms A_k. Let L be the range of the



(a) (b)

Figure 3: Local explanation using SHAP for a single prediction in the Heloc dataset (**a**) XGB, (**b**) and NN model.



(a) (b)

Figure 4: Local explanation using LIME for a single prediction in the Heloc dataset, (**a**) XGB, (**b**) and NN model.

normalized scale (i.e., $L = 1$ for $[0, 1]$).

$$d(A_k, A_l) = \frac{1}{L} \max_{i:x_{il}>x_{ik}} (x_{il} - x_{ik}) \tag{8}$$

If the set of indices where $x_{il} > x_{ik}$ is empty, then $d(A_k, A_l) = 0$.

3. Determine Outranking Relation: An alternative A_k is said to outrank A_l ($A_k \succeq A_l$) if and only if its concordance index meets a minimum threshold $\bar{c}$ and its discordance index does not exceed a maximum threshold $\bar{d}$: $c(A_k, A_l) \geq \bar{c}$ AND $d(A_k, A_l) \leq \bar{d}$. The result of ELECTRE I is an outranking graph. Further analysis of this graph (e.g., identifying the *kernel*, which contains all non-outranked alternatives) provides a final set of preferred explanation methods. Unlike scoring methods that produce a total ranking, ELECTRE I's primary output is this kernel set. Therefore, in our results table, we report only the members of the kernel, as the method does not provide a relative ranking for the outranked alternatives.

3.3.7 Framework Application and Comparative Analysis

The application of our framework, summarized in Algorithms 1, 2, and 3, involves applying these three MCDA methods to the performance scores of the candidate explanation methods. We analyze the rankings produced under various pre-defined weighting scenarios to: (1) identify explanation methods that consistently perform well, (2) understand the sensitivity of the selection to the choice of MCDA technique, and (3) provide practitioners with a transparent tool for selecting an explanation method that best aligns with their requirements.

IF “MSinceMostRecentInqexcl7day” > 0.59 **AND “**NetFractionRevolvingBurden” <= -7.00
AND “PercentTradesNeverDelq”< 97.00 **AND “**ExternalRiskEstimate” > 80
AND “NumTotalTrades” <= 1.78 **AND “**PercentTradesWBalance” <= 67.00
AND“MaxDelq2PublicRecLast12M”<= 6.00**AND “**NumTrades90Ever2DerogPubRec”> 0.00
AND “PercentInstallTrade” ≤ 32.00 **AND “**NumSatisfactoryTrade” <= 27.00
AND “MSinceMostRecentInqexcl7day” <= 11.00 **AND** “MSinceOldestTradeOpen” > 241.00
THEN Prediction: “Bad”
WITH precision = 0.494 **AND** Coverage = 0.062

(a)

IF “PercentInstallTrade” > 0.59 **AND** “PercentTradesWBalance” <= 50.00
AND “NumSatisfactoryTrade” > 13.00 **AND** “MSinceMostRecentInqexcl7day” < -1.00
AND “NetFractionRevolvingBurden”<=58 **AND** “ExternalRiskEstimate” > 75.00
AND “AverageMInFile”<= 96.00 **AND** “NumRevolvingTradesWBalance”<=33.00
AND “NumBank2NatlTradesWHighUtilization”<= − 1.00 **AND** “NumInqLast6M”<=0.00
AND “NumTotalTrades” <= 30.00 **AND** “MaxDelqEve” <= 6.00
THEN Prediction: “Bad”
WITH precision = 0.678 **AND** Coverage = 0.128

(b)

Figure 5: For a single prediction in the Heloc dataset, the local explanations provided by Anchor (**a**) XGB, (**b**) and NN model.

Table 1: Measured explanation evaluation metrics, with average and standard deviation values computed on 50 randomly selected samples from the HELOC dataset. Where ↑ indicates that higher values are better, ↓ indicates that lower values are better, and bolded values are the best values.

XAI	Model	FI (↑)	ST (↓)	ID (↑)	SE (↓)	CO (↑)	FA (↑)
SHAP	**NN**	**1.00 ± 0.00**	0.39 ± 0.01	**0.81 ± 0.01**	0.03 ± 0.01	**0.40 ± 0.02**	0.46 ± 0.01
	XGB	0.96 ± 0.01	0.36 ± 0.02	0.68 ± 0.02	0.03 ± 0.01	0.31 ± 0.02	**0.49 ± 0.01**
	RF	0.94 ± 0.01	0.44 ± 0.01	0.55 ± 0.02	0.05 ± 0.01	0.38 ± 0.01	0.42 ± 0.01
	LR	0.91 ± 0.02	0.61 ± 0.03	0.10 ± 0.03	0.21 ± 0.02	0.10 ± 0.01	0.20 ± 0.01
LIME	**NN**	0.99 ± 0.01	**0.30 ± 0.01**	0.61 ± 0.02	**0.03 ± 0.00**	0.29 ± 0.02	0.47 ± 0.02
	XGB	0.95 ± 0.01	0.40 ± 0.01	0.67 ± 0.02	0.05 ± 0.01	0.10 ± 0.01	0.35 ± 0.02
	RF	0.94 ± 0.01	0.49 ± 0.02	0.28 ± 0.03	0.07 ± 0.01	0.30 ± 0.02	0.29 ± 0.01
	LR	0.89 ± 0.02	0.54 ± 0.02	0.08 ± 0.02	0.38 ± 0.02	0.00 ± 0.00	0.19 ± 0.01
Anchor	**NN**	0.88 ± 0.01	0.50 ± 0.01	0.10 ± 0.02	0.15 ± 0.02	0.10 ± 0.02	0.30 ± 0.02
	XGB	0.82 ± 0.01	0.50 ± 0.01	0.08 ± 0.01	0.24 ± 0.02	0.09 ± 0.01	0.29 ± 0.01
	RF	0.79 ± 0.02	0.56 ± 0.02	0.01 ± 0.02	0.22 ± 0.01	0.10 ± 0.01	0.24 ± 0.01
	LR	0.69 ± 0.02	0.64 ± 0.02	0.00 ± 0.02	0.39 ± 0.02	0.00 ± 0.01	0.22 ± 0.01

4 Experimental Setup

This section discusses the experimental setup and data we used to evaluate our proposed method. We used Scikit-Learn Pedregosa et al. (2011) to train the ML models and TensorFlow-2.4 Sergeev & Del Balso (2018) to train and evaluate our NN model. To explain the Ml/DL models, we used the SHAP Lundberg & Lee (2017), the Anchor Ribeiro et al. (2018), and LIME Ribeiro et al. (2016) library. All models were trained on an Intel Core i9 Processor and 128GB RAM option with NVIDIA GeForce RTX 3090 Ti GPU.

Datasets: To validate the effectiveness of the proposed approach, we use three real-world datasets namely, home equity line of credit (Heloc) Demajo et al. (2020), German Credit Asuncion & Newman (2007), and Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) Angwin et al. (2016). The Heloc dataset is from FICO which consists of information on Heloc applications, with $9,871$ samples for 24 continuous features. We use this dataset to predict whether an applicant made payments without being 90 days overdue. In contrast, the German Credit dataset comprises of 20 features capturing the bank account balance, loan information, demographics, employment information, and credit history of $1,000$ loan applicants. The class label here is a loan applicant's credit risk (e.g., high or low). However, the COMPAS dataset contains 7 features capturing information about the criminal history, demographics, and prison time of $4,937$ defendants. Each defendant in the dataset is labeled either as high-risk or low-risk for recidivism based on the COMPAS algorithm's risk score. We used the 70% samples from these datasets to train the ML/DL models and their remaining 30% for testing.

5 Experimental Results

This section presents the detailed results of our proposed framework for trustworthy explanation selection from a set of explanation methods.

Table 2: Measured explanation evaluation metrics, with average and standard deviation values computed on 50 randomly selected samples from the German Credit dataset. Where ↑ indicates that higher values are better, ↓ indicates that lower values are better, and bolded are the best values.

XAI	Model	FI (↑)	ST (↓)	ID (↑)	SE (↓)	CO (↑)	FA (↑)
SHAP	**NN**	**0.99 ± 0.01**	0.40 ± 0.02	**0.78 ± 0.02**	0.05 ± 0.01	0.34 ± 0.02	0.45 ± 0.02
	XGB	0.93 ± 0.01	**0.39 ± 0.01**	0.69 ± 0.02	**0.04 ± 0.01**	0.32 ± 0.01	0.43 ± 0.01
	RF	0.92 ± 0.01	0.50 ± 0.02	0.48 ± 0.03	0.07 ± 0.01	**0.37 ± 0.02**	0.32 ± 0.01
	LR	0.91 ± 0.01	0.65 ± 0.02	0.01 ± 0.01	0.19 ± 0.01	0.10 ± 0.01	0.15 ± 0.02
LIME	**NN**	0.97 ± 0.01	0.44 ± 0.01	0.70 ± 0.02	0.05 ± 0.01	0.26 ± 0.02	**0.46 ± 0.01**
	XGB	0.94 ± 0.01	0.49 ± 0.02	0.75 ± 0.02	0.05 ± 0.01	0.19 ± 0.02	0.39 ± 0.01
	RF	0.94 ± 0.01	0.54 ± 0.02	0.40 ± 0.03	0.08 ± 0.01	0.27 ± 0.02	0.31 ± 0.02
	LR	0.90 ± 0.01	0.56 ± 0.02	0.10 ± 0.01	0.30 ± 0.02	0.00 ± 0.01	0.15 ± 0.02
Anchor	**NN**	0.84 ± 0.02	0.50 ± 0.01	0.08 ± 0.01	0.17 ± 0.01	0.10 ± 0.02	0.38 ± 0.02
	XGB	0.84 ± 0.01	0.56 ± 0.02	0.01 ± 0.01	0.28 ± 0.02	0.05 ± 0.01	0.25 ± 0.01
	RF	0.75 ± 0.01	0.65 ± 0.02	0.00 ± 0.02	0.25 ± 0.02	0.12 ± 0.01	0.32 ± 0.01
	LR	0.79 ± 0.01	0.71 ± 0.02	0.00 ± 0.01	0.45 ± 0.02	0.00 ± 0.01	0.29 ± 0.01

5.1 ML/DL Model Explanation

The local explanation results using the LIME and SHAP methods for the same data instances from the Heloc dataset for the XGB and NN models are shown in Figures 4 and 3. In Figure 4a, the features such as *MSinceMostRecentDel*, *NetFractionRevolvingBurde*, *AverageMInFil*, *ExternalRiskEstimate* etc., are the ones that contribute to bad class classification, while the features such as *MaxDelqEve*, *NumSatisfactoryTrade*, etc., that have a negative effect on bad class classification. In contrast, in Figure 4b, features such as *MSinceMostRecentDel*, *NumSatisfactoryTrade*, *AverageMInFile*,etc., appear to be the most influential features for bad class classification and features such as *NumRevolvingTradesWBalanc*, *NumTotalTrade*, *NumTrades60Ever2DerogPubRec*, etc., are the most dominating features and have a negative impact in bad class classification for the NN model. Interestingly, the SHAP and LIME explanation methods provide different feature rankings for the same instance in the predicted same class (e.g., bad) explanation, leading to disagreement problems. For example, in Figure 3a, *NumSatisfactoryTrades* is the most important feature for bad class classification in the SHAP explanation, while *MSinceMostRecentDelq* is the most important feature for the same class classification in the LIME explanation, as shown in Figure 4a for the NN model. A similar phenomenon is also observed for the Anchor explanation method, as shown in Figure 5. One thing we notice from these three post-hoc explanation techniques is that they may disagree with each other for the same instance. Indeed, different post-hoc explanation techniques have various goals, leading to an inconsistent and unreliable view of explanation Krishna et al. (2022). Thus, solely looking at the feature ranking does not give sufficient disagreement problems with the predicted class explanation. Therefore, we use six explanation evaluation metrics to measure the intuitions behind the dispute in the predicted class explanation. The next section will evaluate the explanation methods against these metrics.

5.2 Performance of Explanation Methods

This section evaluates SHAP, LIME, and Anchor methods across six explanation quality metrics: fidelity (FI), stability (ST), identity (ID), separability (SE), faithfulness (FA), and consistency (CO). To evaluate and measure the disagreement problem in a principled manner, we randomly select 50 samples from the test datasets.Tables 1, 2, and 3 report the mean and standard deviation of these metrics for the HELOC, German Credit, and COMPAS datasets. It is observed that the SHAP method performs better regarding the NN model most of the time. For instance, SHAP achieves a fidelity score of 1.0 in the NN model, which is $1.02\times$ and $1.14\times$ higher than LIME and Anchor, respectively, indicating relatively low disagreement. However, LIME shows better performance on the COMPAS dataset. Notably, We see that Anchor underperforms in the LR model setting, likely due to the mismatch between sparse rule-based explanations and linear decision boundaries Ribeiro et al. (2018). While for the HELOC dataset, LIME consistently demonstrates higher stability than SHAP across all models. Additionally, the NN model exhibits generally poor stability across all datasets. For example, in the German Credit dataset, SHAP yields a notably lower stability score with the XGB model compared to other ML models, indicating reduced robustness in explanation consistency. However, SHAP consistently shows higher identity scores, suggesting greater consistency for identical inputs. For example, on the HELOC dataset, SHAP achieves an identity score of 0.814 for the NN model, which is $1.34\times$ and $7.90\times$ higher

Table 3: Measured explanation evaluation metrics, with average and standard deviation values computed on 50 randomly selected samples from the Compass dataset. Where ↑ indicates that higher values are better, ↓ indicates that lower values are better, and bolded values are the best values.

XAI	Model	FI (↑)	ST (↓)	ID (↑)	SE (↓)	CO (↑)	FA (↑)
SHAP	**NN**	0.95 ± 0.01	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.78 \pm 0.01}$	$\mathbf{0.04 \pm 0.01}$	$\mathbf{0.35 \pm 0.01}$	$\mathbf{0.42 \pm 0.01}$
	XGB	0.90 ± 0.01	0.35 ± 0.01	0.71 ± 0.01	0.04 ± 0.01	0.31 ± 0.01	0.41 ± 0.01
	RF	0.90 ± 0.01	0.41 ± 0.02	0.41 ± 0.02	0.06 ± 0.01	0.29 ± 0.01	0.37 ± 0.02
	LR	0.89 ± 0.01	0.59 ± 0.02	0.10 ± 0.01	0.31 ± 0.02	0.10 ± 0.01	0.27 ± 0.01
LIME	**NN**	$\mathbf{0.96 \pm 0.01}$	0.41 ± 0.01	0.58 ± 0.02	0.05 ± 0.01	0.22 ± 0.01	0.40 ± 0.01
	XGB	0.95 ± 0.01	0.46 ± 0.02	0.49 ± 0.02	0.09 ± 0.01	0.14 ± 0.01	0.33 ± 0.01
	RF	0.90 ± 0.01	0.51 ± 0.02	0.35 ± 0.02	0.10 ± 0.01	0.20 ± 0.01	0.30 ± 0.01
	LR	0.87 ± 0.01	0.55 ± 0.02	0.01 ± 0.01	0.39 ± 0.02	0.00 ± 0.01	0.15 ± 0.02
Anchor	**NN**	0.82 ± 0.02	0.52 ± 0.02	0.09 ± 0.01	0.21 ± 0.01	0.09 ± 0.01	0.38 ± 0.02
	XGB	0.80 ± 0.01	0.59 ± 0.02	0.01 ± 0.01	0.30 ± 0.01	0.01 ± 0.01	0.29 ± 0.01
	RF	0.80 ± 0.01	0.61 ± 0.02	0.00 ± 0.01	0.27 ± 0.01	0.10 ± 0.01	0.31 ± 0.01
	LR	0.75 ± 0.01	0.69 ± 0.02	0.00 ± 0.00	0.41 ± 0.02	0.00 ± 0.01	0.25 ± 0.01

Table 4: Comprehensive MCDA rankings of the explanation methods for the NN model. The table shows the final rank order across three datasets and seven distinct weighting scenarios, demonstrating the sensitivity of the *"best"* method selection to user priorities. Scores for SAW and TOPSIS are shown in brackets (); for ELECTRE I, the kernel (set of non-outranked alternatives) is shown.

Framework Element		**HELOC Dataset**			**German Credit Dataset**			**COMPAS Dataset**		
Scenario	**MCDA**	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 1**	**Rank 2**	**Rank 3**
Balanced	SAW	SHAP (0.71)	LIME (0.68)	Anchor (0.45)	SHAP (0.64)	LIME (0.62)	Anchor (0.43)	SHAP (0.66)	LIME (0.58)	Anchor (0.48)
	TOPSIS	SHAP (0.83)	LIME (0.68)	Anchor (0.00)	SHAP (0.85)	LIME (0.75)	Anchor (0.00)	SHAP (0.89)	LIME (0.57)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{SHAP}	-	-	{SHAP}	-	-
Fidelity-focused	SAW	SHAP (0.87)	LIME (0.86)	Anchor (0.79)	SHAP (0.86)	LIME (0.85)	Anchor (0.75)	LIME (0.82)	SHAP (0.81)	Anchor (0.73)
	TOPSIS	SHAP (0.97)	LIME (0.91)	Anchor (0.00)	SHAP (0.97)	LIME (0.91)	Anchor (0.00)	LIME (0.96)	SHAP (0.89)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{SHAP}	-	-	{LIME}	-	-
Stability-focused	SAW	LIME (0.72)	SHAP (0.66)	Anchor (0.49)	SHAP (0.65)	LIME (0.62)	Anchor (0.47)	SHAP (0.70)	LIME (0.61)	Anchor (0.50)
	TOPSIS	LIME (0.99)	SHAP (0.61)	Anchor (0.00)	SHAP (0.94)	LIME (0.71)	Anchor (0.00)	SHAP (0.99)	LIME (0.53)	Anchor (0.00)
	ELECTRE I	{LIME, SHAP}	-	-	{SHAP}	-	-	{SHAP}	-	-
Identity-focused	SAW	SHAP (0.78)	LIME (0.68)	Anchor (0.42)	SHAP (0.75)	LIME (0.69)	Anchor (0.41)	SHAP (0.77)	LIME (0.60)	Anchor (0.44)
	TOPSIS	SHAP (1.00)	LIME (0.73)	Anchor (0.00)	SHAP (1.00)	LIME (0.88)	Anchor (0.00)	SHAP (1.00)	LIME (0.66)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{SHAP}	-	-	{SHAP}	-	-
Separability-focused	SAW	SHAP (0.78)	LIME (0.78)	Anchor (0.56)	LIME (0.71)	SHAP (0.71)	Anchor (0.53)	SHAP (0.75)	LIME (0.64)	Anchor (0.55)
	TOPSIS	SHAP (0.90)	LIME (0.90)	Anchor (0.00)	LIME (0.83)	SHAP (0.83)	Anchor (0.00)	SHAP (0.95)	LIME (0.69)	Anchor (0.00)
	ELECTRE I	{LIME, SHAP}	-	-	{LIME, SHAP}	-	-	{SHAP}	-	-
Consistency-focused	SAW	SHAP (0.73)	LIME (0.64)	Anchor (0.43)	SHAP (0.69)	LIME (0.59)	Anchor (0.44)	SHAP (0.71)	LIME (0.55)	Anchor (0.45)
	TOPSIS	SHAP (1.00)	LIME (0.54)	Anchor (0.00)	SHAP (0.99)	LIME (0.70)	Anchor (0.00)	SHAP (1.00)	LIME (0.45)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{SHAP}	-	-	{SHAP}	-	-
Faithfulness-focused	SAW	LIME (0.70)	SHAP (0.69)	Anchor (0.47)	LIME (0.68)	SHAP (0.67)	Anchor (0.49)	SHAP (0.68)	LIME (0.62)	Anchor (0.51)
	TOPSIS	LIME (0.88)	SHAP (0.83)	Anchor (0.00)	LIME (0.95)	SHAP (0.89)	Anchor (0.00)	SHAP (0.87)	LIME (0.74)	Anchor (0.00)
	ELECTRE I	{LIME, SHAP}	-	-	{LIME, SHAP}	-	-	{SHAP}	-	-

than LIME and Anchor, respectively, highlighting significant disparities in identity and stability across the methods. These findings underscore the need for systematic evaluation when selecting explanation techniques.

Furthermore, for the separability metric, LIME generally outperforms SHAP, achieving lower values across most models. For instance, in the HELOC dataset with the NN model, LIME records a separability score of 0.026, which is approximately $1.2\times$ and $5.61\times$ lower than SHAP and Anchor, respectively. However, SHAP achieves lower separability in the COMPAS dataset. Notably, the NN model consistently underperforms across datasets, while the RF model with LIME yields the lowest separability in COMPAS. Regarding consistency, values remain distant from the ideal score of 1.0, highlighting instability and disagreement across explanation methods. Indeed, faithfulness scores are also below ideal. For instance, in the HELOC dataset, SHAP achieves a mean faithfulness of 0.494 for the XGB model, which is $1.40\times$ and $1.70\times$ higher than LIME and Anchor, respectively, indicating that SHAP generally produces more faithful explanations. However, faithfulness varies significantly across methods and models, emphasizing that **no single explanation method consistently yields faithful explanations**. These findings reinforce that explanation methods can yield unstable and inconsistent results across instances, underscoring the core of the *disagreement problem*.

5.3 Comparative MCDA Ranking and Analysis

In the preceding section, we observed that no explanation method consistently provides precise explanations, and no singular explanation evaluation metric can comprehensively evaluate explanation methods. Therefore, we are combining the evaluation metrics of explanation methods by employing the MCDA method for selecting the proper explanation method from a set of post-hoc explanation techniques for a given set of instances. The results of our comparative MCDA framework are presented in Tables 4 and 5 across the NN and XGB models, revealing a rich and multifaceted landscape of explanation method performance. By analyzing the rankings across seven distinct weighting scenarios and three datasets for the NN model, we can move beyond simplistic best explanation method claims to a more nuanced, context-aware understanding. Our analysis yields three primary insights. First, **SHAP emerges as a consistently robust and well-rounded choice.** For instance, in the **balanced** scenario, SHAP is the unambiguous winner across all three datasets and all three MCDA methods. This dominance extends to priority scenarios focused on explanation integrity: in the **identity-focused** and **consistency-focused** scenarios, SHAP is again the clear Rank 1 choice, as shown in Table 4. This strong, cross-logic agreement suggests that for general-purpose use or when the reliability and consistency of explanations are paramount, SHAP provides a powerful and justifiable default selection.

Second, our framework clearly demonstrates that **the optimal explanation method is highly sensitive to both user priorities and the underlying dataset.** For instance, in the **fidelity-focused** scenario, while SHAP wins on the HELOC and German Credit datasets, LIME emerges as the top choice for COMPAS. An even starker shift occurs in the **stability-focused** scenario, where LIME is the top-ranked method for HELOC, but SHAP wins decisively on both German Credit and COMPAS. These results empirically invalidate any one-size-fits-all heuristic (e.g., *"always use LIME for stability"*) and underscore the critical need for a data-driven, context-specific evaluation like our proposed framework provides.

Third, and most critically, the comparison between MCDA methods reveals **the crucial role of decision logic in exposing hidden trade-offs.** This is most evident in scenarios where LIME is the top-ranked method by the compensatory SAW and TOPSIS techniques. For instance, in the **stability-focused** scenario on the HELOC dataset, and the **faithfulness-focused** scenarios on the HELOC and German Credit, while SAW/TOPSIS ranks LIME first, the non-compensatory ELECTRE I method finds LIME and SHAP to be incomparable (i.e., both are in the kernel). A deeper analysis reveals why: LIME's superior score on the single prioritized metric is offset by significant weaknesses on other criteria (such as its poor Identity score), as shown in Table 1. ELECTRE's discordance condition flags this trade-off, preventing LIME from being declared a clear winner, an insight completely invisible to the other methods. This demonstrates that our framework not only ranks alternatives but also makes the very nature of the decision, simple compensation versus a risk-averse choice with trade-offs, transparent to the user. Note that, unlike scoring methods that produce a total ranking, ELECTRE I's primary output is this kernel set. Therefore, we report only the members of the kernel in the Rank 1 column, as the method does not provide a relative ranking for the outranked alternatives, as shown in Table 4. Across all scenarios, the Anchor method is consistently ranked last, suggesting its rule-based approach is less suited for these complex, non-linear models compared to its counterparts. Thus, these results demonstrate that our framework not only selects the most suitable explanation method from a given set of techniques for specific instances but also facilitates a transparent, context-aware decision-making process. A similar phenomenon is also observed for identity, separability, consistency, and faithfulness-focused scenarios across all datasets. It is also noted that, we observed similar trend for the XGB model across all datasets, as shown in Table 5.

6 Discussion

The experimental results underscore a fundamental shift in how the XAI disagreement problem can be approached. Instead of searching for a single "best" explanation method, our comparative MCDA framework provides a more pragmatic and defensible path: selecting the "most suitable" method for a specific context and set of priorities. This shift carries key implications for practitioners and researchers, which we elaborate on below.

Implications for Practitioners: For ML practitioners and domain experts, this work provides a transparent, structured methodology to navigate the complex landscape of XAI tools. Rather than relying on anecdotal evidence or the popularity of a method, users can articulate their priorities (e.g., *"for this regulatory audit, explanation stability is paramount"*) and receive a well-justified recommendation. The comparison between compensatory (SAW/TOPSIS) and non-compensatory (ELECTRE I) methods is especially valuable, as it highlights the distinction between a *"best on*

Table 5: Comprehensive MCDA rankings of the explanation methods for the XGBoost model. The table shows the final rank order across three datasets and seven distinct weighting scenarios, demonstrating the sensitivity of the *"best"* method selection to user priorities. Scores for SAW and TOPSIS are shown in brackets (); for ELECTRE I, the kernel (set of non-outranked alternatives) is shown.

Framework Element		HELOC Dataset			German Credit Dataset			COMPAS Dataset		
Scenario	**MCDA Method**	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 1**	**Rank 2**	**Rank 3**
Balanced	SAW	SHAP (0.68)	LIME (0.60)	Anchor (0.42)	SHAP (0.66)	LIME (0.62)	Anchor (0.39)	SHAP (0.66)	LIME (0.56)	Anchor (0.37)
	TOPSIS	SHAP (1.00)	LIME (0.78)	Anchor (0.00)	SHAP (0.92)	LIME (0.89)	Anchor (0.00)	SHAP (0.95)	LIME (0.70)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{LIME, SHAP}	-	-	{SHAP}	-	-
Fidelity-Focused	SAW	SHAP (0.84)	LIME (0.82)	Anchor (0.70)	LIME (0.83)	SHAP (0.82)	Anchor (0.72)	LIME (0.80)	SHAP (0.79)	Anchor (0.67)
	TOPSIS	SHAP (0.97)	LIME (0.87)	Anchor (0.00)	LIME (0.96)	SHAP (0.91)	Anchor (0.00)	LIME (0.98)	SHAP (0.87)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{LIME}	-	-	{LIME}	-	-
Stability-Focused	SAW	SHAP (0.69)	LIME (0.64)	Anchor (0.46)	SHAP (0.67)	LIME (0.61)	Anchor (0.42)	SHAP (0.68)	LIME (0.58)	Anchor (0.39)
	TOPSIS	SHAP (0.95)	LIME (0.87)	Anchor (0.00)	SHAP (1.00)	LIME (0.63)	Anchor (0.00)	SHAP (1.00)	LIME (0.62)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{SHAP}	-	-	{SHAP}	-	-
Identity-Focused	SAW	SHAP (0.71)	LIME (0.65)	Anchor (0.39)	LIME (0.70)	SHAP (0.69)	Anchor (0.36)	SHAP (0.71)	LIME (0.56)	Anchor (0.34)
	TOPSIS	SHAP (1.00)	LIME (0.98)	Anchor (0.00)	LIME (1.00)	SHAP (0.92)	Anchor (0.00)	SHAP (1.00)	LIME (0.61)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{LIME}	-	-	{SHAP}	-	-
Separability-Focused	SAW	SHAP (0.85)	LIME (0.80)	Anchor (0.64)	SHAP (0.83)	LIME (0.82)	Anchor (0.60)	SHAP (0.83)	LIME (0.77)	Anchor (0.62)
	TOPSIS	SHAP (1.00)	LIME (0.90)	Anchor (0.00)	SHAP (0.98)	LIME (0.97)	Anchor (0.00)	SHAP (1.00)	LIME (0.88)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{LIME, SHAP}	-	-	{SHAP}	-	-
Consistency-Focused	SAW	SHAP (0.72)	LIME (0.56)	Anchor (0.41)	SHAP (0.69)	LIME (0.58)	Anchor (0.36)	SHAP (0.70)	LIME (0.53)	Anchor (0.33)
	TOPSIS	SHAP (1.00)	LIME (0.13)	Anchor (0.00)	SHAP (1.00)	LIME (0.45)	Anchor (0.00)	SHAP (1.00)	LIME (0.34)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{SHAP}	-	-	{SHAP}	-	-
Faithfulness-Focused	SAW	SHAP (0.71)	LIME (0.61)	Anchor (0.45)	SHAP (0.68)	LIME (0.62)	Anchor (0.41)	SHAP (0.68)	LIME (0.58)	Anchor (0.39)
	TOPSIS	SHAP (1.00)	LIME (0.47)	Anchor (0.00)	SHAP (0.93)	LIME (0.82)	Anchor (0.00)	SHAP (0.98)	LIME (0.76)	Anchor (0.00)
	ELECTRE I	{SHAP}	-	-	{SHAP}	-	-	{SHAP}	-	-

average" choice and one that avoids critical, "veto-level" weaknesses. For instance, let us recall the healthcare scenario discussed in Section 1, our framework allows a practitioner to evaluate whether SHAP or LIME provides more reliable insights for clinicians based on specific diagnostic priorities. By formalizing this selection, our framework helps ensure the chosen explanations are consistent and trustworthy, fostering collaborative decision-making between the AI system and healthcare professionals and mitigating trust issues.

Methodological Implications for Researchers: Our work demonstrates the value of applying established decision science principles to challenges in XAI. The discordance observed between compensatory (e.g., SAW, TOPSIS) and non-compensatory (ELECTRE I) methods highlights that the choice of aggregation logic itself is a crucial, often overlooked, aspect of meta-evaluations in AI. It suggests that future XAI evaluation benchmarks should consider offering a suite of aggregation techniques rather than a single, fixed leaderboard score. Furthermore, for researchers developing novel explanation methods, our framework provides a multi-faceted benchmark. Instead of claiming superiority on a single metric, new methods can be evaluated on their performance profiles across various weighting scenarios, offering a more complete and honest assessment of their strengths and trade-offs against existing techniques. This provides a structured approach to identify where a new method is susceptible to disagreement and how it compares to established alternatives.

7 Limitations and Future Work

Although our proposed method addresses a key challenge for the practical adoption of XAI, it has few limitations. First, we evaluated our framework using three ML models and one relatively simple DL model. With the growing interest in foundation models and generative AI, we have not yet applied our framework to more complex DL architectures, such as Transformer-based networks or diffusion models. Extending our evaluation to these models is an important direction for future work. Second, our study considered three post-hoc explanation methods and six explanation evaluation metrics. While this set is representative, it is not exhaustive and may not capture all relevant aspects of explanation quality. Third, we focused solely on feature-importance-based explanation methods in this work. As part of our future research, we plan to evaluate our proposed method alongside other explanation families, including example-based, counterfactual, and visual methods. Moreover, the weighting scenarios used in our experiments were pre-defined for illustrative purposes. Real-world deployments would benefit from a more formal preference elicitation process involving stakeholders, such as the Analytic Hierarchy Process (AHP) Saaty (2008). Finally, our framework currently resolves disagreements at the *method-selection* level; it does not merge or reconcile feature-level disagreements for a single prediction, which remains an open avenue for future investigation.

8 Conclusion

This paper addressed the critical *disagreement problem* in XAI by proposing and validating a novel evaluation framework based on comparative Multi-Criteria Decision Analysis. By leveraging SAW, TOPSIS, and ELECTRE I, we moved beyond a single aggregated score to a transparent, preference-driven process for selecting the most suitable post-hoc explanation method. Validated on four ML/DL models, three explanation methods, and three real-world datasets, our framework successfully navigates the trade-offs between competing explanation quality criteria (e.g., fidelity and stability), highlighting not only the best-performing methods under different scenarios but also the critical sensitivities in the decision logic itself. This work provides practitioners and researchers with a robust, adaptable tool to make more principled and justifiable choices about explainability, thereby fostering greater trust and reliability in AI systems.

References

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799, 2022.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica (2016). *URL: https://www. propublica. org/article/machine-bias-risk-asses sments-in-criminal-sentencing*, 2016.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

Brian Barr, Noah Fatsi, Leif Hancox-Li, Peter Richter, Daniel Proano, and Caleb Mok. The disagreement problem in faithfulness metrics. *arXiv preprint arXiv:2311.07763*, 2023.

Szymon Bobek, Paweł Bałaga, and Grzegorz J Nalepa. Towards model-agnostic ensemble explanations. In *Computational Science–ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part IV*, pp. 39–51. Springer, 2021.

Longbing Cao. Ai in finance: A review. *Available at SSRN 3647625*, 2020.

Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H Bach, and Himabindu Lakkaraju. 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*, pp. 203–214, 2022.

Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

Lara Marie Demajo, Vince Vella, and Alexiei Dingli. Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*, 2020.

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

Sina Fazelpour and Will Fleisher. The value of disagreement in ai design, evaluation, and alignment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2138–2150, 2025.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3681–3688, 2019.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems*, 35:5256–5268, 2022.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020.

Lukas Klein, Carsten Lüth, Udo Schlegel, Till Bungert, Mennatallah El-Assady, and Paul Jäger. Navigating the maze of explainable ai: A systematic approach to evaluating methods and metrics. *Advances in Neural Information Processing Systems*, 37:67106–67146, 2024.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Gabriel Laberge, Yann Batiste Pequignot, Mario Marchand, and Foutse Khomh. Tackling the xai disagreement problem with regional explanations. In *International Conference on Artificial Intelligence and Statistics*, pp. 2017–2025. PMLR, 2024.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. *arXiv preprint arXiv:2106.12543*, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42, 2022.

Michael Neely, Stefan F Schouten, Maurits JR Bleeker, and Ana Lucic. Order in the court: Explainable ai methods prone to disagreement. *arXiv preprint arXiv:2105.03287*, 2021.

Enea Parimbelli, Tommaso Mario Buonocore, Giovanna Nicora, Wojtek Michalowski, Szymon Wilk, and Riccardo Bellazzi. Why did ai get this one wrong?—tree-based explanations of machine learning model predictions. *Artificial Intelligence in Medicine*, 135:102471, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Zhongang Qi, Saeed Khorram, and Fuxin Li. Visualizing deep networks by optimizing with integrated gradients. In *CVPR Workshops*, volume 2, pp. 1–4, 2019.

Omer Reingold, Judy Hanwen Shen, and Aditi Talati. Dissenting explanations: Leveraging disagreement to reduce model overreliance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21537–21544, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 444–448. IEEE, 2022.

Thomas L Saaty. Decision making with the analytic hierarchy process. *International journal of services sciences*, 1 (1):83–98, 2008.

Vera Schmitt, Luis-Felipe Villa-Arenas, NIls Feldhus, Joachim Meyer, Robert P Spang, and Sebastian Möller. The role of explainability in collaborative human-ai disinformation detection. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 2157–2174, 2024.

Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson. Reckoning with the disagreement problem: Explanation consensus as a training objective. *arXiv preprint arXiv:2303.13299*, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Alexander Sergeev and Mike Del Balso. Horovod: fast and easy distributed deep learning in tensorflow. *arXiv preprint arXiv:1802.05799*, 2018.

Priscylla Silva, Vitoria Guardieiro, Brian Barr, Claudio Silva, and Luis Gustavo Nonato. Visagreement: Visualizing and exploring explanations (dis) agreement. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34:9391–9404, 2021.

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

David Solís-Martín, Juan Galán-Páez, and Joaquín Borrego-Díaz. On the soundness of xai in prognostics and health management (phm). *Information*, 14(5):256, 2023.

Vinitra Swamy, Jibril Frej, and Tanja Käser. The future of human-centric explainable artificial intelligence (xai) is not post-hoc explanations. *Journal of Artificial Intelligence Research*, 84, 2025.

Ziqi Zhao, Yucheng Shi, Shushan Wu, Fan Yang, Wenzhan Song, and Ninghao Liu. Interpretation of time-series deep models: A survey. *arXiv preprint arXiv:2305.14582*, 2023.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.