Unifying VXAI: A Systematic Review and Framework for the Evaluation of Explainable AI

Anonymous authors
Paper under double-blind review

Abstract

Modern AI systems frequently rely on opaque black-box models, most notably Deep Neural Networks, whose performance stems from complex architectures with millions of learned parameters. While powerful, their complexity poses a major challenge to trustworthiness, particularly due to a lack of transparency. Explainable AI (XAI) addresses this issue by providing human-understandable explanations of model behavior. However, to ensure their usefulness and trustworthiness, such explanations must be rigorously evaluated. Despite the growing number of XAI methods, the field lacks standardized evaluation protocols and consensus on appropriate metrics. To address this gap, we conduct a systematic literature review following the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* guidelines and introduce a unified framework for the *eValuation of XAI (VXAI)*. We identify 362 relevant publications and aggregate their contributions into 41 functionally similar metric groups. In addition, we propose a three-dimensional categorization scheme spanning explanation type, evaluation contextuality, and explanation quality desiderata. Our framework provides the most comprehensive and structured overview of VXAI to date. It supports systematic metric selection, promotes comparability across methods, and offers a flexible foundation for future extensions.

1 Introduction

Explainable AI (XAI) is a research area of growing interest to both AI researchers and practitioners. It aims to alleviate the black-box issue of current deep-learning models, which can reach stunning performances at the expense of their interpretability (Vilone & Longo, 2021). Government-affiliated initiatives, such as the European Union High-Level Expert Group on AI (2019), the U.S. National Institute of Standards and Technology (2023), and the DARPA initiative (Gunning & Aha, 2019), identified XAI as a crucial part of Trustworthy AI. Especially as it helps AI systems in serving the "right to explain" its decisions (Goodman & Flaxman, 2017) and fosters user trust through understanding of the system (Morandini et al., 2023). XAI already plays a fundamental role in making high-stakes AI systems more trustworthy (Saarela & Podgorelec, 2024; Xua & Yang, 2024), with broad applications in areas such as healthcare, finance, autonomous driving, natural disaster detection, energy management, military and remote sensing (Adadi & Berrada, 2018; Markus et al., 2021; Saraswat et al., 2022; Kadir et al., 2023; Hosain et al., 2024; Höhl et al., 2024). Furthermore, explainability is used to help with other dimensions of trustworthiness like privacy, robustness, or fairness (Doshi-Velez & Kim, 2017; Yang et al., 2019; Arrieta et al., 2020; Das & Rad, 2020; Markus et al., 2021; Rawal et al., 2021; Agarwal et al., 2022b).

However, XAI is not a silver bullet. Van der Waa et al. (2021) point out, that humans tend to trust predictions more readily when an explanation is provided, often without carefully examining the explanation itself. This lack of critical scrutiny can lead to unwarranted trust, especially when decisions are taken based on incorrect or misleading explanations (Eiband et al., 2019; Jesus et al., 2021). To make matters worse, different XAI algorithms may result in conflicting explanations for the same model and sample (Krishna et al., 2022). Therefore, simply providing *any* explanation is not sufficient, but it is important to assess the quality of the explanation at hand (Sovrano et al., 2021). Unfortunately, while there is a plethora of XAI methods, evaluation of explanations is still an immature research area (Ribera & Lapedriza, 2019), with many studies

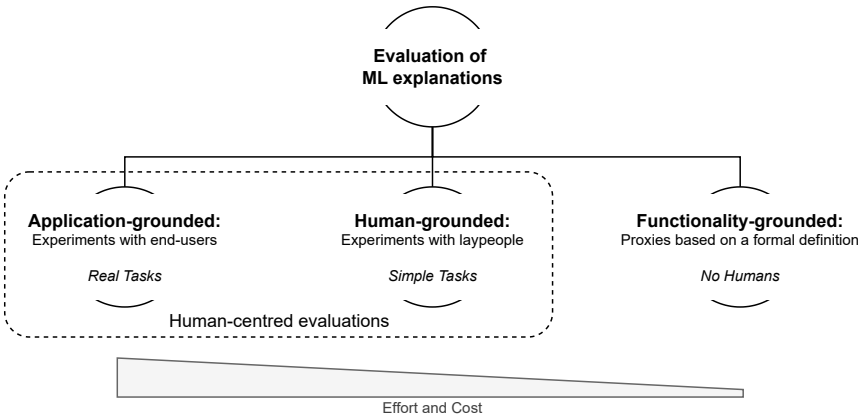


Figure 1: The classification of XAI evaluation into human-grounded and functionality-grounded evaluation, adapted from the classification framework by Doshi-Velez & Kim (2017) and its visualization by Zhou et al. (2021).

relying on the notion of a good explanation as "You'll know it when you see it", providing anecdotal evidence (i.e. small-scale qualitative validation) (Doshi-Velez & Kim, 2017; Nauta et al., 2023; Saarela & Podgorelec, 2024). Especially in computer vision tasks, evaluation through qualitative inspection of a few examples can be appealing (Ibrahim & Shafiq, 2023). However, unstructured qualitative examination leads to highly subjective results, as humans struggle at judging the value of XAI explanations (Adebayo et al., 2018; Buçinca et al., 2020; Hase & Bansal, 2020). In addition, such evaluations risk cherry-picking favorable examples and offer no reliable foundation for comparing different explanation methods across studies or practitioners. For the same given explanation, human ratings vary depending on both the task itself (Franklin & Lagnado, 2022) and the participant's cultural background (Peters & Carman, 2024). Evaluation is further complicated by the lack of ground-truth for the explanations, as this would require knowledge about the model's internal reasoning process (Samek et al., 2019; Markus et al., 2021; Samek et al., 2021; Bommer et al., 2024; Ortigossa et al., 2024).

Because evaluation is still not performed consistently and seldom systematically (Adadi & Berrada, 2018; Lipton, 2018; Payrovnaziri et al., 2020; Messina et al., 2022; Lopes et al., 2022; De Camargo et al., 2023; Kadir et al., 2023; Nauta et al., 2023; Mohamed et al., 2024; Naveed et al., 2024; Saarela & Podgorelec, 2024; Salih et al., 2024a), the community frequently calls to develop comprehensive and unified evaluation standards (Pinto & Paquette, 2024; Saarela & Podgorelec, 2024; Xua & Yang, 2024). A central motivation behind such efforts is to enable the comparison of explanations and asses whether explainability is achieved (Markus et al., 2021; Zhou et al., 2021)

One of the most prevalent taxonomies reported in the literature (Vilone & Longo, 2021; Zhou et al., 2021; Elkhawaga et al., 2023), and illustrated in Figure 1, is the distinction proposed by Doshi-Velez & Kim (2017) between human-grounded and functionality-grounded evaluation methods. The former includes qualitative and quantitative evaluations by laypeople and experts, while the latter consists of (semi-)automatic metrics.

Since explanations are meant to aid humans, human-grounded evaluation remains the gold standard to assess their effectiveness in assisting humans (Doshi-Velez & Kim, 2017; Gunning & Aha, 2019; Miller et al., 2017). However, the faithfulness (i.e., technical correctness) of an explanation and the plausibility to humans do not necessarily correlate (Wiegreffe & Pinter, 2019; Jacovi & Goldberg, 2020; Atanasova, 2024a). Therefore, human-grounded evaluation of an explanation's comprehensibility should be distinguished from functionality-grounded evaluation of its faithfulness (Nauta et al., 2023). Especially, humans can't confidently attribute whether an unexpected explanans (i.e., the information provided to explain a decision[1]) is caused by a faulty explanation (process[1]) or a flawed black-box model (Robnik-Šikonja & Bohanec, 2018; Zhang et al., 2019a); see Figure 2 for an illustration. In both cases, the consequences can be severe, either reducing trust in

[1]The exact definitions of explanandum, explanation, and explanans are provided at the end of Section 1.

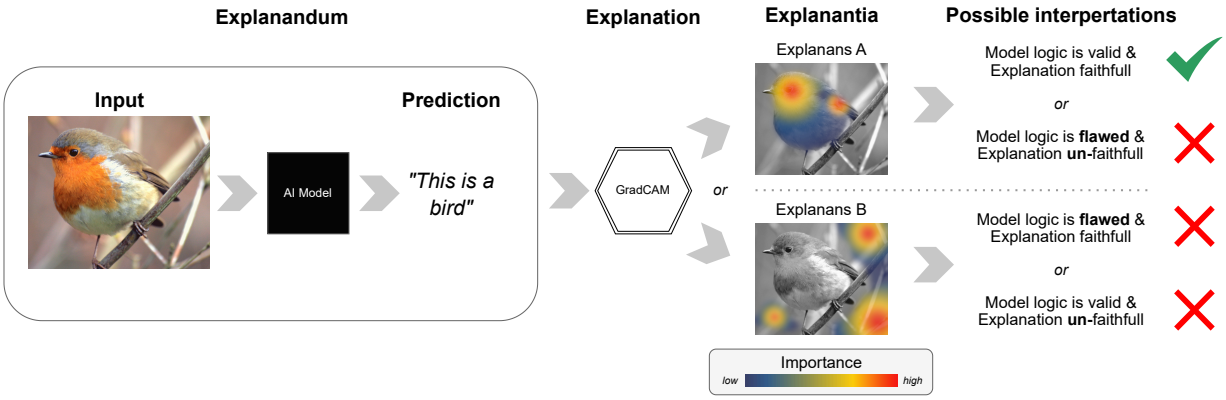


Figure 2: The heatmaps show two alternative example explanantia[1], indicating which input regions were deemed decisive for the model's decision (the explanandum[1]). A qualitative inspection allows for multiple interpretations, as it is unclear whether a) both the model and the explanation process (explanation[1]) are correct or flawed (top), or b) one is correct and the other failed (bottom).

a well-functioning model or, more critically, reinforcing trust in a flawed one. Further, human evaluation, especially through the system's developers, is prone to confirmation bias (Doshi-Velez & Kim, 2017; Lipton, 2018)

There exist a number of surveys and guidelines that address human-centered evaluations (Hoffman et al., 2018; Miller, 2019; Chromik & Schuessler, 2020; Holzinger et al., 2020; Franklin & Lagnado, 2022; Hsiao et al., 2021; Jesus et al., 2021; Langer et al., 2021; Mohseni et al., 2021; van der Waa et al., 2021; Silva et al., 2023). Unfortunately, the range of reviews dedicated to functionality-grounded evaluation is still limited. This is despite the advantage of offering objective, quantitative metrics without requiring human experiments, which can save both time and cost (Doshi-Velez & Kim, 2017; Samek et al., 2019; Zhou et al., 2021). Most existing studies are narrow and restricted to a specific application domain (Giuste et al., 2022; Arreche et al., 2024), including cybersecurity (Pawlicki et al., 2024), medical image classification (Patrício et al., 2023; Chaddad et al., 2024), electronic health record data (Payrovnaziri et al., 2020), data and knowledge engineering (Li et al., 2020b), or time-series classification (Theissler et al., 2022). Others focus on particular XAI approaches, such as visual explanations in CNNs (Mohamed et al., 2022) or instance-based explanations (Bayrak & Bach, 2024). Moreover, many surveys dedicate only limited attention to evaluation metrics, with their primary focus placed on the XAI methods themselves (Carvalho et al., 2019; Ding et al., 2022; Minh et al., 2022; Mohamed et al., 2022; Ali et al., 2023; Clement et al., 2023; Patrício et al., 2023; Chaddad et al., 2024; Gongane et al., 2024; Xua & Yang, 2024). By contrast, this review focuses exclusively on functionality-grounded evaluation across domains and is applicable to a wide range of XAI approaches.

Contributions

Despite the growing number of proposed metrics, a comprehensive and unified framework for functionality-grounded evaluation is still missing. Further, the inconsistent use of terms such as interpretability, comprehensibility, understandability, transparency, and explainability (Koh & Liang, 2017; Guidotti et al., 2018; Arrieta et al., 2020; Markus et al., 2021) hampers conceptual clarity and comparability across approaches. To address this gap, we introduce a framework called **eValuation of Explainable Artificial Intelligence (VXAI)**, aimed at unifying functionality-grounded evaluation for XAI.
Our contributions are as follows:

- We perform a systematic literature review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines by Page et al. (2021), identifying 362 relevant publications that introduce or utilize evaluation metrics.
- We aggregate these into 41 functionally similar metric groups, capturing common methodological patterns across the literature.

- We propose a three-dimensional categorization scheme consisting of desiderata, explanation type, and evaluation contextuality, and use it to organize the identified metrics.
- To our knowledge, this results in the most comprehensive and unified VXAI framework to date and provides an extensible foundation for future research.

The remainder of this review is structured as follows: In Section 2, we first present related studies on the topic of VXAI to motivate the need for this systematic review, before introducing the desiderata of XAI in Section 3. Further, Section 4 introduces our research method used to search for and identify relevant metrics. The categorization scheme and results are presented in Section 5, where we introduce our categorization framework (Subsection 5.1) and summarize the identified metrics (Subsection 5.2), complemented by a visual overview in Figure 6. A deeper discussion of these findings is provided in Section 6, while comprehensive descriptions of the metrics alongside references are listed in Appendix B. We conclude the review in Section 7, discussing the results and future paths for the area of VXAI.

Terminology

To avoid ambiguous language, throughout the paper we stick to the terminology of the XAI Handbook by Palacio et al. (2021): The goal of XAI is to facilitate understanding by providing insights into an *explanandum* ("What is to be explained"), usually a model or a model's decision. To accomplish this, we leverage an *explanation*, which is the process of getting insight into the explanandum. The resulting output of this process is the *explanans*, which provides the user with information about the model's inner workings. In a mathematical sense, the explanation can be viewed as a function that maps an explanandum to an explanans. For example, the explanandum could be a CNN's classification of a given input image. The explanation might be an algorithm such as GradCAM (Selvaraju et al., 2017), and the resulting heatmap is the explanans, which highlights important features. We use the Latin plural forms explananda (explanandum) and explanantia (explanans) throughout. When we refer to VXAI, we include both the evaluation of the method (explanation) and its output (explanans), since most evaluation metrics necessarily assess the quality of explanations through the quality of their generated outputs.

As defined by the XAI Handbook, *interpretation* (or *interpretability*) refers to the subsequent assignment of meaning to an explanation. It describes the process through which a human infers knowledge about the explanandum using the explanans. This step significantly influences the success of the explanation and also depends on the receiving human's (the *explainee*'s) mental model.

Terminology

explanandum (pl. **explananda**): What is to be explained, i.e. a model and its prediction.
explanation (pl. **explanations**): The process of explaining, i.e. the XAI algorithm.
explanans (pl. **explanantia**): The explaining information, i.e. the output of an explanation.

2 Related Work

Although the field of XAI has gained popularity over the past years, there is still no extensive and unified evaluation framework for XAI metrics. Various surveys have explored XAI and VXAI from different angles, ranging from human-grounded evaluation to technical metrics. Table 1 gives an overview over 30 such XAI reviews from the past years.

While evaluation of XAI is frequently given less attention in XAI surveys, 23 of these reviews directly focus on the topic of VXAI. Besides functionality-grounded evaluation, a second school of thought is concerned with human-grounded evaluation of explanations through qualitative expert evaluations or quantitative user studies, with representative surveys for this domain available as well (Sokol & Flach, 2020; Rawal et al., 2021; Naveed et al., 2024). Nevertheless, a considerable number of 19 reports focus specifically on the topic of functionality-grounded evaluation. Unfortunately, most of these surveys focus on a subset of well-known metrics, whereas only 14 surveys gathered VXAI metrics in a systematic or semi-systematic literature review.

Work	VXAI Focus	Functionality-Grounded VXAI	(Semi-)Systematic	Date ↓	Desiderata	Limited to	Reported Metrics
This work	✓	✓	✓	Jan 2025	Parsimony, Plausibility, Coverage, Fidelity, Continuity, Consistency, Efficiency		41 metrics from 362 sources
Klein et al.	✓	✓		Jan 2025	Faithfulness, Robustness, Complexity	Feature Attributions; Computer Vision	20 metrics
Pawlicki et al.	✓	✓	✓	Oct 2024		Cybersecurity	86 metrics
Awal & Roy	✓	✓		Jun 2024	Reliability, Consistency	Rule Explanations	6 metrics
Bayrak & Bach	✓	✓	✓	Apr 2024		Counterfactuals	66 metrics
Bommer et al.	✓	✓		Mar 2024	Robustness, Faithfulness, Complexity, Localization, Randomization	Climate Science	10 metrics
Li et al.	✓	✓		Dec 2023	Faithfulness	Feature Attributions	6 metrics
Alangari et al.	✓			Aug 2023	Correctness, Comprehensibility, Stability		59 metrics
Le et al.	✓	✓	✓	Aug 2023	Co-12†		86 metrics from 17 toolkits
Salih et al.	✓		✓	Aug 2023		Cardiology	27 metrics
Kadir et al.	✓	✓	✓	Jul 2023			80 metrics
Hedström et al.	✓	✓		Apr 2023	Faithfulness, Robustness, Localization, Complexity, Axiomatic, Randomization	Feature Attribution	27 metrics
Schwalbe & Finzel			✓	Jan 2023			11 metrics (already grouped)
Agarwal et al.	✓	✓		Nov 2022	Faithfulness, Stability	Feature Attributions	11 metrics
Coroama & Groza	✓			Nov 2022			26 metrics
Verma et al.		✓		Nov 2022		Counterfactuals	9 metrics (already grouped)
Belaid et al.	✓	✓		Oct 2022	Fidelity, Fragility, Stability, Simplicity, Stress, Other	Feature Attributions	22 metrics
Cugny et al.	✓	✓		Oct 2022			6 metrics
Lopes et al.	✓		✓	Aug 2022	Fidelity (Completeness, Soundness), Interpretability, Broadness, Simplicity, Clarity)		43 metrics
Yuan et al.		✓	✓	Jul 2022	Fidelity, Sparsity, Stability, Accuracy	Graph Neural Networks	7 metrics
Löfström et al.	✓		✓	Mar 2022			10 metrics
Vilone & Longo	✓		✓	Dec 2021			36 metrics
Bodria et al.		✓	✓	Nov 2021			6 metrics
Sovrano et al.	✓			Oct 2021	Similarity, Exactness, Fruitfulness		22 metrics
Ras et al.				Sep 2021			*13 sources in text* (metrics not listed)
Mohseni et al.			✓	Aug 2021	Fidelity, Trustworthiness		15 metrics
Yeh & Ravikumar	✓	✓		Jun 2021			7 metrics
Nauta et al.	✓	✓	✓	May 2021	Co-12†		28 metrics (already grouped)
Zhou et al.	✓	✓		Jan 2021	Fidelity (Completeness, Soundness), Interpretability, Broadness, Simplicity, Clarity)		17 metrics
Samek & Müller				Sep 2019			*16 sources in text* (metrics not listed)
Yang et al.	✓	✓	✓	Aug 2019	Generalizability, Fidelity, Persuasibility		*40 sources in text* (metrics not listed)

†Co-12: Correctness, Output-Completeness, Consistency, Continuity, Contrastivity, Covariate Complexity, Compactness, Composition, Confidence, Context, Coherence, Controllability

Table 1: Overview of recent XAI reviews, sorted by date. The table indicates whether each survey primarily focused on evaluation metrics, whether it reported mainly functionality-grounded metrics, and whether a (semi-)structured review was conducted. The date refers to the earliest available point in the article's timeline; either the database query, submission, or publication, depending on what was reported. For each survey, we also report the desiderata used to classify the metrics and any limitations regarding explanation type or application domain. Note that not all surveys systematically listed their assessed metrics, so the reported metric count may vary depending on the method of extraction.

Further, numerous of the referenced reviews either lack an extensive list of desiderata and focus only on a subset of them, or limit their research to specific types of explanations[2] or application domains.

There are five reviews, that we consider most similar to this work, as they present systematic functionality-grounded VXAI surveys. Le et al. (2023) and Nauta et al. (2023) both categorize the identified metrics based on a scheme of 12 properties, namely the Co-12 framework, which we discuss in more detail in Section 3. However, Le et al. (2023) restrict their analysis to metrics available through public XAI or VXAI toolkits (e.g., Quantus (Hedström et al., 2023)) and do not report on metrics introduced in the literature but not implemented in such libraries. Although there is some overlap with the study by Nauta et al. (2023), particularly in the inclusion of some identical metrics, their review also incorporates studies that merely apply VXAI metrics rather than introducing them. In contrast, our work provides detailed descriptions and categorizations of each identified metric. The review by Kadir et al. (2023) covers a broad range of domains and explanation types[2] and reports a wide variety of metrics. It also groups several metrics by method, a strategy shared by our work. However, it does not adopt a categorization scheme based on the desiderata fulfilled by individual metrics. Notably, the recent reviews from Bayrak & Bach (2024) and Pawlicki et al. (2024) report a high number of individual metrics for VXAI. However, both limit the scope of their review considerably, either in terms of application domain (Pawlicki et al., 2024) or explanation type[2] (Bayrak & Bach, 2024). In contrast, our work includes all metrics reported to date and introduces a categorization scheme based on explicitly defined desiderata (see Section 3). Finally, many of the metrics we identified were introduced only recently, underlining the need for this more recent literature review.

While previous reviews report between 10 and 90 individual metrics, our work introduces a unified structure by aggregating over 360 individual metrics into 41 conceptually related groups. This enables clearer comparison and interpretation across metrics. Unlike most surveys, we do not limit our analysis to specific explanation types or application domains, ensuring broader applicability across the XAI landscape.

3 Desiderata of XAI

A well-founded evaluation of XAI methods requires clearly defined criteria for what constitutes a good explanation. To establish such criteria, we must first reflect on the role of explanations in the context of XAI. According to the definition in the XAI Handbook (Palacio et al., 2021), explaining a model and its behavior is a two-stage process: first, factual information about the model's decision process is generated (the explanans); this is then interpreted by the human user. The first stage can be evaluated using technical criteria that assess whether the model's reasoning has been captured truthfully and reliably. The second stage depends on the interpretability of the explanation, which can be assessed using general cognitive principles, even in the absence of a specific user model.

To capture the multifaceted nature of explanation quality, a number of desiderata have been proposed in the literature. We interpret these as functionality-grounded expectations that reflect the demands of both stages of the explanation process. In this section, we first provide an overview of existing desiderata proposed in prior work. We then introduce a unified framework that systematically describes the requirements for ensuring technical soundness and for bridging the interpretation gap.

3.1 Common Formulation of Desiderata

Several XAI surveys report that there is no ubiquitous consensus on appropriate desiderata, with some of the categories related to goals pursued *through* XAI, rather than standalone desiderata of XAI, e.g., Trustworthiness, Acceptance, or Fairness (Doshi-Velez & Kim, 2017; Langer et al., 2021; Vilone & Longo, 2021; Elkhawaga et al., 2023). Hence, we conduct a scoping review, reporting the main desiderata used by different authors and analyzing the similarities as well as differences in their formulations. For the sake of brevity, we exclude some of the papers listed in Table 1, as the missing ones either overlap considerably (e.g. Awal & Roy (2024) and Klein et al. (2024)), rely on a different notion of desiderata (e.g., Sovrano

[2]The *explanation type* refers to both the design of the explanation algorithm and, consequently, the nature of the resulting explanans, as introduced in Subsection 5.1.2.

et al. (2021)), or use no desiderata at all. We will first present these frameworks using the authors' original terminology before introducing our own categorization scheme.

The famous **Co-12** properties, introduced by Nauta et al. (2023) and reused by Le et al. (2023), constitute one of the most extensive existing frameworks for categorizing XAI metrics. They group the properties along three different dimensions: Content (*Correctness, Completeness, Consistency, Continuity, Contrastivity, Covariate Complexity*), Presentation (*Compactness, Composition, Confidence*), and User (*Context, Coherence, Controllability*). While the first dimension focuses on the information contained in the explanans, the second and third dimensions address the way the information is conveyed. Although some of these human-centered properties can be measured through proxies, others may mainly be evaluated through human-grounded evaluation.

Zhou et al. (2021), based on the taxonomy of Markus et al. (2021), define *Interpretability* and *Fidelity* as the two major components of explainability. The former is concerned with providing understandable explanantia and includes the properties of *Clarity*, *Broadness*, and *Parsimony*. Fidelity, on the other hand, refers to how accurately an explanans reflects the model's behavior, and consists of the properties of *Completeness* and *Soundness*.

The framework proposed by Robnik-Šikonja & Bohanec (2018), which was adopted by Carvalho et al. (2019) and Molnar (2020), differentiates between properties of explanations and individual explanantia. Investigating the properties of methods (i.e. explanations), they consider *Translucency*, *Portability*, and *Algorithmic Complexity*, which can all be interpreted as desiderata, while *Expressive Power* is a descriptive property. The properties of individual explanantia include *Comprehensibility*, *Importance*, *Representativeness*, *Fidelity*, and *Stability*. However, their categorization encompasses further properties, which we do not consider as proper desiderata of XAI: *Accuracy*, *Novelty*, *Certainty*, and *Consistency*. Accuracy is a measurement of the underlying black-box model, while Novelty and Certainty are rather explanantia themselves, than properties of general explanations. Further, Consistency between different black-box models is not necessarily a useful measure, as different models may derive similar predictions based on different reasoning (see Rashomon Effect (Breiman, 2001; Leventi-Peetz & Weber, 2022)).

The famous XAI review by Guidotti et al. (2018), inspired by earlier works such as Andrews et al. (1995) and Johansson et al. (2004), reports three less fine-grained desiderata: *Interpretability*, *Fidelity*, and *Accuracy*. Similar to previously discussed reviews, Interpretability describes human understandability, while Fidelity measures how well the explanans imitates the black box, and Accuracy focuses on predictive performance, which is outside the scope of XAI in our context. Additionally, *Consistency* is introduced by Andrews et al. (1995), expecting reproducible explanations, while Johansson et al. (2004) emphasize the explanation's algorithmic *Scalability* and *Generality*.

Alvarez-Melis & Jaakkola (2018b), Jesus et al. (2021), and Alangari et al. (2023a) all report a similar set of desiderata. The understandability of explanations is measured in terms such as *Interpretability*, while *Faithfulness* and the corresponding desiderata give insight into how truthful the explanation is to the underlying black-box model. All three works further report the *Stability* of explanations as a desired property, assessing whether explanantia on similar inputs are similar.

In their Quantus toolkit, Hedström et al. (2023) (and the follow-up study by Bommer et al. (2024) as well), categorize their metrics partly through desiderata, namely *Faithfulness*, *Robustness*, and *Complexity*. Simultaneously, part of their metrics are grouped by their conceptual similarity, including *Localization*, *Randomization*, and *Axiomatic*.

Finally, the Compare-xAI benchmark by Belaid et al. (2022) organizes functional tests into six categories, namely *Fidelity*, the robustness-related *Stability* and *Fragility*, the interpretability desideratum *Simplicity*, and the explanation-methods-focused *Stress* and *Portability* (which they integrate under "Other").

While many existing frameworks overlap conceptually, a unified and practically usable categorization scheme for VXAI metrics is still lacking. This requires a structured set of desiderata that defines what makes a good explanation and supports consistent classification of metrics. Prior work often enforces a rigid one-to-one mapping between metrics and desiderata; in contrast, we decouple these dimensions, defining a set of mostly independent desiderata to which each metric may contribute individually or jointly. Lightweight frameworks

tend to omit critical aspects of explanation quality, while broader ones sometimes include goals that are not intrinsic to the explanation itself (e.g., accuracy). We restrict our scope to properties that reflect the explanation rather than the underlying model and clarify excluded cases after presenting our set. Although all desiderata rely on proxies, we limit ourselves to properties that are quantifiable in principle. Highly abstract or vague notions lacking empirical grounding are omitted. Lastly, our framework is designed to be extensible, allowing the integration of future desiderata as the field evolves.

3.2 Proposed Framework of Desiderata

Building on the desiderata frameworks established above and our findings on VXAI metrics, we propose a set of seven desiderata to serve as a categorization scheme for VXAI.

Our goal is to offer a principled yet practical structure that enables consistent classification while avoiding the limitations of prior frameworks. These are either too narrow to accommodate relevant metrics or too broad and include properties beyond explainability. While properties such as fairness are often measured using XAI methods, we consider them beyond the scope of VXAI, because they assess the model's behavior rather than the explanation itself.

Building on the two-stage view of explaining described by Palacio et al. (2021) (i.e., presenting factual information followed by human interpretation), we define two complementary dimensions of explanation quality. The **Technical** (T) dimension comprises desiderata that assess the factual correctness, robustness, and reliability of the explanation, ensuring that it faithfully reflects the model's reasoning. In contrast, the **Interpretability** (I) dimension captures how the explanation is conveyed and how accessible, intuitive, and useful it is to a general-purpose user. This separation is aligned with existing frameworks such as the Co-12 properties (Nauta et al., 2023) and the taxonomy by Zhou et al. (2021). The desiderata are designed to be as independent from each other as possible, allowing for reliable quantification of different aspects relevant to trustworthy XAI.

We present our categorization scheme and its relation to other frameworks in Figure 3 and introduce them in more detail in the following paragraphs. In total, we define seven desiderata, two associated with Interpretability, and five belonging to the Technical dimension:

(I) **Parsimony**: The explanation should keep the explanans concise to support interpretability.
(I) **Plausibility**: The explanation should shape the explanans to align with human expectations.
(T) **Coverage**: The explanation should provide an explanans for every explanandum.
(T) **Fidelity**: The explanation should make the explanans reflect the model's true reasoning.
(T) **Continuity**: The explanation should ensure that similar explananda yield similar explanantia.
(T) **Consistency**: The explanation should produce stable explanantia across repeated evaluations.
(T) **Efficiency**: The explanation should compute the explanans efficiently and broadly.

3.2.1 Parsimony

The explanation should keep the explanans concise to support interpretability.

The primary purpose of an explanation is to convey information about the black-box model or its decision process to humans. Therefore, the resulting explanans must be expressed in a way that the human mind can grasp easily to increase the explanations success. While actual interpretability, can only be evaluated through human-grounded evaluation, Parsimony is one of the most prevalent proxies defined to serve as a functionality-grounded approximation. Since our mental capacity is limited and we tend to struggle with an overload of information (Miller, 1956; 2019; Alangari et al., 2023a), providing short and simple explanantia helps humans understand more effectively.

Nauta et al. (2023) introduce the property of Compactness, arguing that a briefer explanans is easier to understand. Similarly, they use Covariate Complexity to assess how complex the features are that constitute the explanans, where higher interpretability is supported by providing a few high-level concepts in favor of a very granular explanans. Both of these aspects are summarized under Parsimony by Markus et al. (2021) and

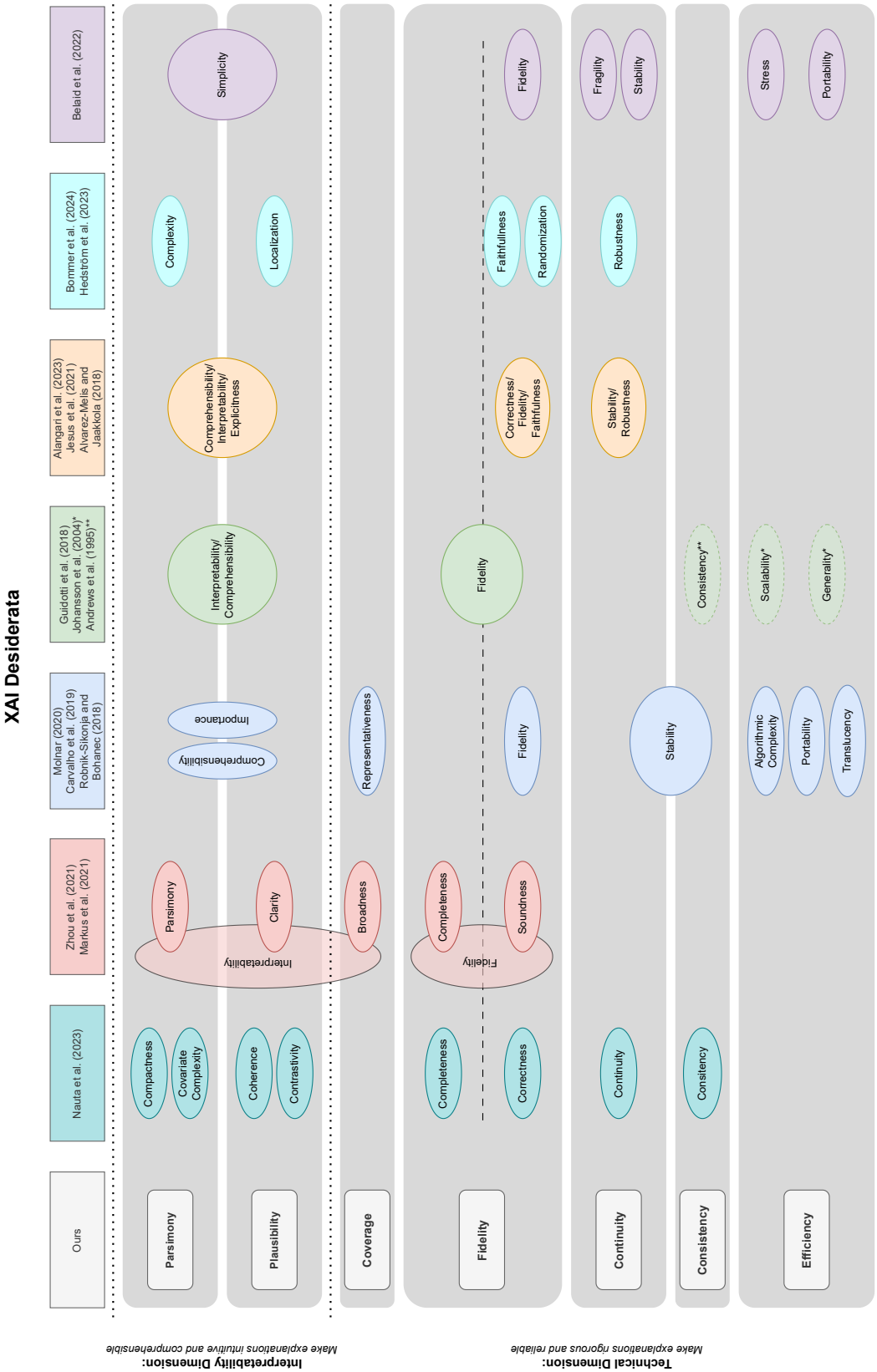


Figure 3: Comparison of our defined desiderata with those used in related work. Desiderata not included in our framework are omitted. Desiderata reported by only individual sources (per column) are marked with an asterisk (*)

Zhou et al. (2021), preferring simpler explanantia over longer or more complex ones. The scheme used in the Quantus library (Hedström et al., 2023; Bommer et al., 2024) defines a group called Complexity. It specifically tests for concise explanantia and aims to have as few features as possible to be easier to understand. The associated interpretability desiderata from other authors are defined less explicitly, but similarly favoring simpler explanantia (Andrews et al., 1995; Johansson et al., 2004; Guidotti et al., 2018), proposing that simple explanantia should be short (Alvarez-Melis & Jaakkola, 2018b; Jesus et al., 2021; Alangari et al., 2023a), promoting small explanantia and only focussing on relevant parts (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020), and expecting an explanans with concentrated information to facilitate human understanding (Belaid et al., 2022).

Following the proposed definitions, we include Parsimony as one of our desiderata. It expects explanantia to be as brief and concise as possible, to ensure that rationales can be understood easily and fast. We focus Parsimony exclusively on this aspect, as other associated properties are either covered by separate desiderata (such as truthfulness of the explanation) or excluded entirely as they are not functionality-grounded (general understandability of the explanation).

3.2.2 Plausibility

The explanation should shape the explanans to align with human expectations.

To improve the acceptance of explanations and facilitate their interpretation, the Plausibility desideratum is contained in most authors' interpretability desiderata. Coherence as defined by Nauta et al. (2023) is the accordance of an explanans with the user's previous knowledge and expectations. The metrics classified under Localization by Hedström et al. (2023) evaluate whether an explanation shows a rationale similar to what humans would expect. This is concordant with the definition of Comprehensibility (Alvarez-Melis & Jaakkola, 2018b; Jesus et al., 2021; Alangari et al., 2023a), which states that an explanans should be similar to what a human expert would choose as the correct rationale. Furthermore, Nauta et al. (2023) introduce Contrastivity, which supports Plausibility, as an explanans should be specific to the given explanandum. Similarly, Clarity is introduced by Markus et al. (2021) and Zhou et al. (2021) expecting explanantia to be unambiguous.

We include the Plausibility desideratum, which encompasses the idea that explanations should align with human knowledge and intuition. On one hand, this includes human expectations towards the result (explanans), e.g., "The model focuses on what a human would focus on". On the other hand, the XAI methods' behavior (explanation) should also be aligned with human intuition, e.g., "The outputs for individual inputs should differ".

3.2.3 Coverage

The explanation should provide an explanans for every explanandum.

The extent to which an explanation or explanans can be applied is considered by two frameworks. Unfortunately, definitions from both surveys are vague. Markus et al. (2021) and Zhou et al. (2021) define the Broadness of an explanation as "how generally applicable" it is, without further elaboration on the implications of this definition. More concretely, Representativeness is presented by Robnik-Šikonja & Bohanec (2018) Carvalho et al. (2019), and Molnar (2020). It reflects the number of explananda that are covered by an individual explanans, although this definition focuses mainly on the distinction between global and local explanation methods.

To add more clarity to these definitions, we include Coverage with an alternative definition. It defines the amount of explananda that are covered by the explanation, i.e. reflecting whether there exists an explanans for every data input or output.

3.2.4 Fidelity

The explanation should make the explanans reflect the model's true reasoning.

Fidelity is one of the most frequently discussed concepts in the literature and combines two closely related aspects: Correctness and Completeness. While some works introduce these as separate desiderata, others group them together under the umbrella of Fidelity.

Correctness refers to whether the explanation truthfully represents the internal logic and decision process of the black-box model. It is one of the most frequently emphasized desiderata across the reviewed frameworks. Without correctness, even the most interpretable or simple explanation may provide no meaningful insight. Terms like Faithfulness, Truthfulness, and Fidelity are often used interchangeably in literature to describe this idea. Correctness encompasses both local fidelity for individual explanantia and global alignment across the dataset (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020). The general consensus is that an explanation should reveal what truly drives the model's outputs (Alvarez-Melis & Jaakkola, 2018b; Markus et al., 2021; Zhou et al., 2021; Belaid et al., 2022; Alangari et al., 2023a; Nauta et al., 2023). It is commonly assessed by how well the explanation reflects or mimics the model's behavior (Andrews et al., 1995; Johansson et al., 2004; Guidotti et al., 2018).

Completeness, in contrast, describes how much of the model's reasoning is captured by the explanation. According to the Co-12 properties by Nauta et al. (2023), an explanation should ideally include the full scope of the model's rationale. Some authors treat Completeness as a sub-aspect of Fidelity (Markus et al., 2021; Zhou et al., 2021), while others define Fidelity itself as the capacity to capture all of the information embodied in the model (Andrews et al., 1995; Johansson et al., 2004; Guidotti et al., 2018).

Although it is theoretically possible to have an explanation that is partially correct but incomplete (e.g., providing a heatmap that highlights only one of several relevant features), or complete but partially incorrect (e.g., including all the right features alongside irrelevant ones), neither scenario is desirable. If key features are missing or irrelevant ones are included, the explanans ultimately misrepresents the model's behavior. While Correctness and Completeness can be distinguished conceptually, they are tightly interwoven in practice and difficult to evaluate in isolation. Since our desiderata are intended to capture orthogonal evaluation dimensions, and these two cannot be meaningfully disentangled, we combine them under the unified criterion of Fidelity.

3.2.5 Continuity

The explanation should ensure that similar explananda yield similar explanantia.

Just as the robustness or stability of a standard AI model is of great interest, similar expectations apply to explainability. Most frameworks highlight this desideratum, and various metrics have been proposed to assess how stable or reliable explanations are. However, the terminology used in literature is inconsistent, at times overlapping and at other times diverging in meaning.

Nauta et al. (2023) introduce the term Continuity as the smoothness of the explanation, i.e. similar explananda should yield similar explanantia. Others refer to this idea as Stability (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020), describing it as the resilience against slight variations in input features that do not alter the model's prediction. The term Robustness is used by Alvarez-Melis & Jaakkola (2018b), Jesus et al. (2021), and Alangari et al. (2023a), to describe the same behavior, referring to it as a key requirement for trustworthy XAI.

The Quantus toolkit reflects the prevalence of this concept, providing a "Robustness" metric category (Hedström et al., 2023; Bommer et al., 2024), which assesses the similarity of explanantia under minor changes in input. Finally, Belaid et al. (2022) cover the same idea under the term Stability. In addition, they assess Fragility, which they define as the resilience of explanations against malicious manipulation, such as adversarial attacks.

Our Continuity desideratum covers both of the mentioned properties. It includes the smoothness of explanations with respect to "naïve" changes in the explanandum that ideally do not affect the model's behavior, as well as the resilience of explanations against malicious manipulation attempts. Note, that this includes changes over the input data as well as the model. We decided to adopt the term Continuity instead of Stability or Robustness, to reduce the possible confusion with model robustness.

3.2.6 Consistency

The explanation should produce stable explanantia across repeated evaluations.

While Continuity investigates the smoothness and similarity between similar but different inputs, the Consistency of explanations for identical inputs also needs to be considered. However, few authors explicitly consider this desideratum.

Consistency is introduced by Nauta et al. (2023) as a direct measure of the determinism of an XAI algorithm. Similarly, one part of the definition of Stability by Robnik-Šikonja & Bohanec (2018), Carvalho et al. (2019), and Molnar (2020) considers variations in explanations based on non-determinism. The oldest formulation of Consistency is given by Andrews et al. (1995) and considers explanation methods to be consistent when they produce equivalent results under repetition.

However, several frameworks additionally consider the similarity of explanantia generated from different models trained on the same data (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020; Nauta et al., 2023). Yet, different models can produce the same prediction while relying on entirely different internal reasoning. This is especially true as there are often multiple valid reasons for the same event, also known as the *Rashomon Effect* (Breiman, 2001; Leventi-Peetz & Weber, 2022).

We include Consistency using the initial formulations, i.e., explanations should be deterministic or self-consistent, always presenting the same explanans for identical explananda. While the latter definition is present in one of the identified metrics, we do not explicitly add it to the definition of our Consistency desideratum, as we do not believe that different explananda (inputs), i.e. different models, necessarily result in identical explanantia (outputs).

3.2.7 Efficiency

The explanation should compute the explanans efficiently and broadly.

Finally, out of practical considerations, we want explanations to be conveniently applicable. This includes considering the range of models or situations in which the algorithm can be effectively applied. Simultaneously, it also includes the time it takes to compute an individual explanans.

The first property is introduced as Portability and Translucency throughout literature (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020). Portability is the variety of models for which an explanation can be used, while Translucency is the necessity of the explanation algorithm to have access to the internals of the model. Similarly, Johansson et al. (2004) measure Generality, given by the restrictions or overhead necessary to apply an explanation to specific models. Belaid et al. (2022) refer to Portability as the diverse set of models to which the explanation can be applied.

Secondly, the Algorithmic Complexity (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020) considers the time it takes to generate an explanans. Naturally, the amount of necessary time depends not only on the inherent complexity of the explanation algorithm, but also on the Scalability, i.e. its ability to efficiently handle larger models and input spaces (Johansson et al., 2004). Using the "Stress test", Belaid et al. (2022) explicitly evaluate the runtime behavior with respect to increasing input size.

We subsume both of these aspects under a general desideratum called Efficiency. It includes the algorithmic or computational properties of the explanation, which might influence the choice of a specific XAI algorithm over another.

3.2.8 Excluded Desiderata

Several of the desiderata introduced in the referenced frameworks were not included in our catalog. We will briefly present each of these and justify why they were excluded.

In the Co-12 properties for XAI, Nauta et al. (2023) introduce **Controllability**, which we exclude for two reasons. First, it measures the extent to which a user can interact with the explanans, which can be relevant in given applications, but is a property of the *presentation* and not necessarily the explanans itself. Second, the only measurement they provide is through user interaction, which we want to avoid in this study, as we focus on functionality-grounded metrics. Their **Composition** property is not measurable either, as it describes the explanans' presentation format. Similarly, **Expressive Power** (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020) describes the format of the explanans and is therefore not included.

The **Context** property (Nauta et al., 2023) describes how relevant a given explanans is to a user. This is mostly covered in the Plausibility desideratum, although their definition is more focused on the needs of specific users and hence is very subjective and not measurable through proxies.

Further, both **Confidence** (Nauta et al., 2023) and the closely related **Certainty** (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020) are excluded. They both describe whether confidence scores of the model's decisions or the explanans are displayed. However, this is not a general desideratum for explanations, but can rather be seen as providing assisting information, which is also an explanans in itself, and is then subject to all the desiderata of our framework. Related to this, **Novelty** (Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020) refers to whether data instances lie within or outside the training distribution. While relevant to assessing model behavior or uncertainty, this property does not characterize the quality of the explanation itself. Instead, it serves as contextual information that may inform or accompany an explanans, but is not an intrinsic requirement for explainability.

As discussed in Subsection 3.2.6, we do not include some of the aspects of **Consistency** as defined by Robnik-Šikonja & Bohanec (2018), Carvalho et al. (2019), and Molnar (2020), who expect several models to generate similar explanantia. This is rooted in the assumption that different models should all follow the same reasoning, which we reject. In contrast, Anders et al. (2020) showed that two models exhibiting the same external behavior, such as predictions, can have radically different inner workings.

While it is a standard metric during model training, **Accuracy** of explanations is also reported by several frameworks (Andrews et al., 1995; Johansson et al., 2004; Guidotti et al., 2018; Robnik-Šikonja & Bohanec, 2018; Carvalho et al., 2019; Molnar, 2020). Importantly, this is only meaningful for specific types of explanations, such as white-box surrogate models (e.g., decision trees, see Subsection 5.1.2 for a definition). It is different from "surrogate fidelity", since this measures the performance against the black-box model's predictions and not the ground-truth. We restrict our framework to explanantia derived from black-box models, rather than evaluating interpretable models that are directly trained on the data as predictive models themselves.

Lastly, the Quantus toolkit (Hedström et al., 2023; Bommer et al., 2024) includes a category labeled as **Axiomatic**, which groups metrics based on their functionality. Specifically, it refers to whether they test for formal properties. This categorization reflects the underlying mechanism of the metric rather than a specific quality criterion. Accordingly, we reassign these metrics to the corresponding desiderata they evaluate.

3.3 Considerations

We want to end the presentation of our desiderata and the resulting classification scheme with a few final considerations.

First, all seven of our desiderata are introduced as independent dimensions. However, in practice, they can not always be separated entirely, which is reflected in some metrics being associated with multiple desiderata. Especially, where a single criterion in other authors' frameworks covers multiple of our desiderata, an inherent connection is given. This can be seen especially in Consistency and Continuity, which both measure some sort of robustness. Conversely, some of our desiderata subsume multiple sub-aspects themselves. This applies especially to Fidelity, which is comprised of Correctness and Completeness.

Second, a truly useful explanation can only be achieved when considering the interplay between the individual Desiderata. Otherwise, trivial explanantia can be constructed to optimize individual desiderata, as we illustrate through the following two examples: a) A Saliency Map, which highlights the entire input image, is certainly complete, as it encompasses all relevant features; however, it is hardly parsimonious and unlikely to be correct, as not every region is going to be truly decisive for the prediction. b) A Concept explanation which identifies the same identical concept regardless of the input may be maximally consistent, continuous, and parsimonious; nevertheless, it will unlikely be either correct or complete, as it disregards input-specific variation.

Further, to ensure the usefulness of XAI to humans, we believe that some hierarchy exists between our defined desiderata. While an exact ordering is implausible, as it depends on the exact setting and needs, at least some general guidelines can be given: **Fidelity** is the foundation of the entire evaluation process. If there is no necessity for an explanation to be correct and complete, it can be completely arbitrary and will not necessarily give any true insights into the underlying model. This would undermine the entire purpose of XAI. If important parts of the rationale are excluded in the explanans, it may result in an incomplete (and hence incorrect) understanding of the decision process. While not as crucial, but still important, **Consistency** and **Continuity** considerably contribute to the trustworthiness of XAI. Without them, explanations may vary unpredictably, as explanantia can change without reasons rooted in the model's behavior. Given these four desiderata, the explanation can be expected to give solid results from a technical viewpoint. Hence, the interpretability dimensions, **Parsimony** and **Plausibility**, can be considered next, as they facilitate understanding, but do not assess whether an explanation is actually reliable. Finally, **Coverage** and **Efficiency** are useful properties of an explanation, but can be seen as a bonus, making them readily applicable to a wide set of explananda, including data and models. However, depending on the use case and context, individual desiderata may be weighted differently or even disregarded entirely; particularly when some aspects are less relevant for a specific application or domain.

Finally, the framework is based on the works presented by other authors, as well as the metrics we identified through our systematic survey. While we consider it extensive for now, it is possible to expand it horizontally in the future. Depending on the requirements, further desiderata may be added. We do not rigorously map each metric to a single desideratum, but instead list several desiderata it contributes to. Therefore, our proposed categorization can be readily extended with further desiderata.

4 Method

In our review, we aim to systematically collect and classify all functionality-grounded metrics relevant to evaluating explanations in the context of XAI. We base our review on the PRISMA (Page et al., 2021) guidelines to make the process transparent. In this section, we describe the entire process used to identify VXAI metrics from the literature. Figure 4 gives an overview of our process.

Preliminary Consideration

The research was preceded by two key observations for the design of the study: First, searching for general XAI terms is not feasible, as the database is intractable (searching Google Scholar for "XAI" gives over 200 thousand results). Second, VXAI metrics are usually introduced alongside an XAI method rather than in a dedicated publication, and there exists no unified vocabulary, making it hard to identify relevant sources from keyword searches in titles or abstracts alone. Hence, we decided to split our research into two stages. First, we performed a database search using a limited set of search terms to build an initial body of potentially relevant sources. In the second phase, we expanded this set through recursive backward snowballing by reviewing the reference lists of the already identified publications.

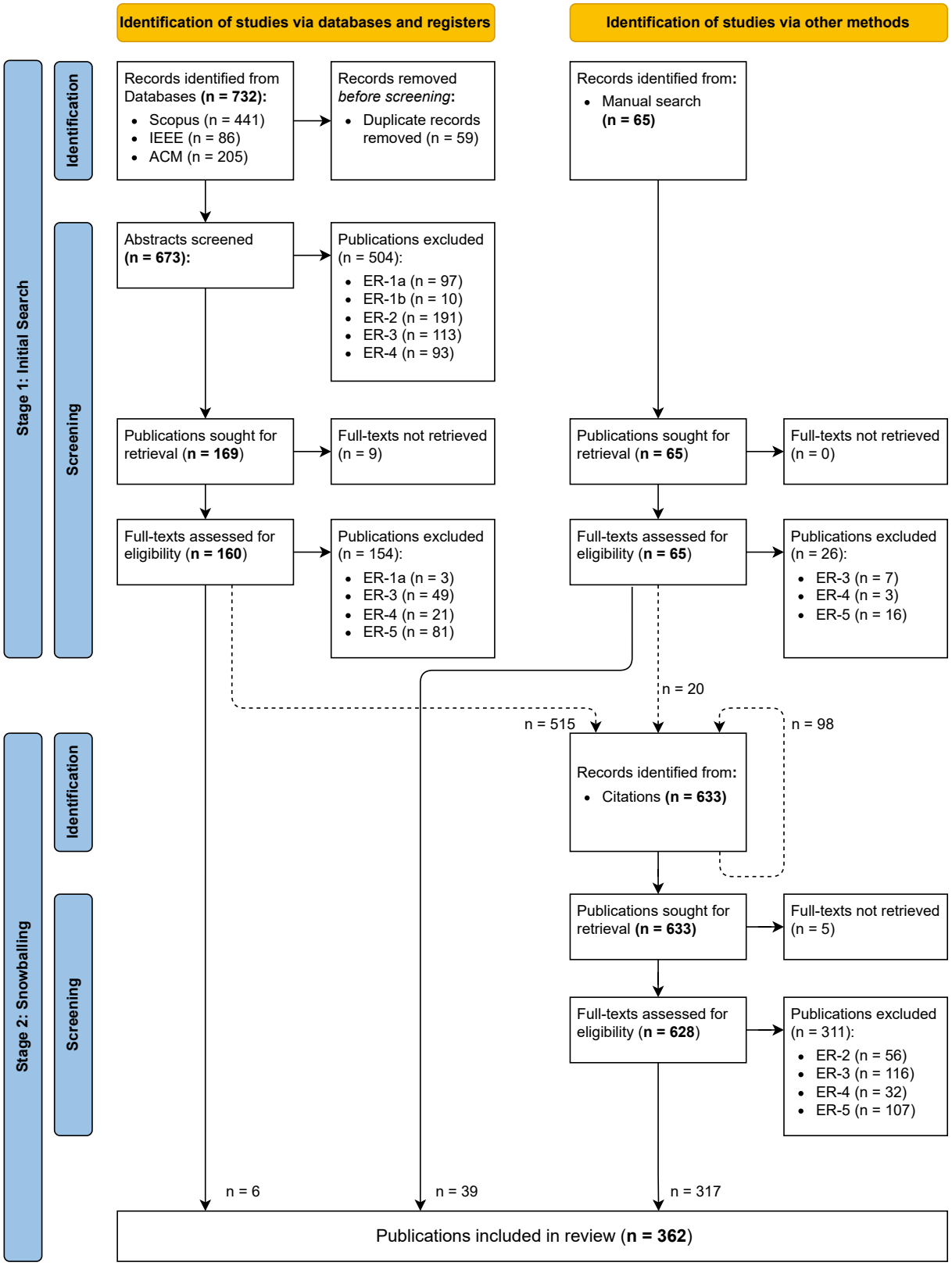


Figure 4: Our search strategy building upon the PRSIMA guidelines (Page et al., 2021). Notably, we split the process into an initial database search phase and a snowballing phase, identifying the most relevant literature in the second phase.

Stage 1: Initial Search

We started with a database meta-search using **Scopus**[3], **IEEE**[4] and **ACM**[5], to identify existing reviews and surveys that point towards further evaluation metrics. Using the advanced search features of each database, we designed research queries based on the following key terms:

```
[Explain* | Interpret*]
          ×[6]
[AI | Artificial Intelligence | ML | Machine Learning]
          AND
[Evaluation | Metric | Quantification]
          AND
[Survey | Review]
```

The exact search string for every database is given in Appendix A. The last search was conducted on January 15, 2025. This resulted in a total of 673 identified articles after de-duplication. We first screened titles and abstracts, excluding sources that were not research articles, did not involve XAI or VXAI, or focused on human-based evaluation. Through this first screening, we excluded 504 articles, using the following exclusion criteria:

- ER-1: General Issues
  - ER-1a: Not a research article
  - ER-1b: Not english
- ER-2: No XAI (or VXAI) at all, or with explainability outside our scope (e.g. ante-hoc data analysis)
- ER-3: Contains XAI, but no systematic evaluation of XAI (e.g. "anecdotal evidence")
- ER-4: Contains systematic VXAI, but only human-centered methods

Out of the 169 articles that passed the initial title and abstract screening, 9 could not be retrieved. For the remaining 160 articles, we conducted full-text screening, applying the exclusion criteria described above. To avoid an intractable number of sources with limited added value, we additionally excluded works that used functionality-grounded VXAI metrics introduced by other articles, introducing the following exclusion criterion:

- ER-5: Contains functionality-grounded evaluation, but does not introduce a new metric itself.

Instead, we added the cited metric to our corpus for the snowballing phase, resulting in 515 potentially relevant sources. Of the articles retrieved though the initial database search, only 6 were ultimately included in our review.

We followed a similar procedure for 65 additional articles identified through other sources, including Google Scholar and personal databases. After excluding 26 of these, 39 articles were directly included in our review. An additional 20 references from these articles were added to the corpus of the snowballing phase.

In total, during the first stage, we included 45 articles in our review (6 from database search and 39 from additional sources). We also compiled 535 additional references for the second-stage backward snowballing (515 from the database search and 20 from the additional sources).

Stage 2: Snowballing

Starting with the corpus of 535 references identified in the first stage, we conducted full-text screening on all records as described above. During the recursive backward snowballing process, we identified and assessed additional 98 references for eligibility. We observed that the set of relevant records quickly began to converge:

[3] https://www.scopus.com/search/form.uri?display=advanced
[4] https://ieeexplore.ieee.org/search/advanced/command
[5] https://dl.acm.org/search/
[6] The × denotes a logical combination of terms.

most newly encountered papers that did not present original metrics (ER-5) cited articles already included in our corpus. We did not perform forward snowballing (i.e., identifying articles that cite our included works), as our focus was on original metric proposals. Including such follow-up papers would have significantly increased the number of records beyond a manageable scope.

In total, we assessed 628 articles during the second phase (535 from the first stage and 98 from recursive snowballing). Out of these we included 317. Compared to the 20% inclusion rate in the first stage, the nearly 50% inclusion rate in the second phase confirms the effectiveness of our strategy. The initial database search primarily yielded secondary sources (e.g., XAI and VXAI reviews), which helped identify original metric proposals during snowballing.

In total, we reviewed 1459 articles, screened 866 in full, and included 362 that originally proposed a VXAI metric or one of its variants.

5 Results

In this section, we present the main results of our literature search. We identified a total of 362 individual references, some of which introduce or modify multiple metrics simultaneously. We sorted and aggregated these metrics based on their methodology, grouping together those that are functionally similar and measure the same properties. Each such group is referred to as an aggregated metric For each aggregated metric, we assessed the associated desiderata, the metric's contextuality, and the suitable explanation types as we define in our categorization scheme below. Having introduced the seven desiderata in Section 3, we now formalize the remaining dimensions of the categorization scheme, before giving an overview of the aggregated metrics. However, to ensure readability, this section will mainly focus on meta-results and statistics of the metrics. Table 2 gives an overview over the metrics, while detailed descriptions with the associated references are presented in Appendix B. A browsable version of the framework will be made available at `https://placeholder.link/available-upon-acceptance`.

5.1 Categorization Scheme

5.1.1 Overview

To facilitate the selection of evaluation metrics for testing explanations, we propose a categorization scheme that groups the identified metrics into functionally similar approaches. Serving both as a conceptual overview and a practical guide, it is structured along three orthogonal dimensions:

(i) the **desiderata** each metric contributes to,
(ii) the degree of dependency on the model and data (**contextuality**), and
(iii) the suitable **explanation types** the metric applies to.

Desiderata: One contribution of this survey is to introduce an extensive and unifying framework for desiderata in the context of VXAI (see Section 3 for details). We apply this framework to all identified metrics. Specifically, for each metric we annotate to which of the seven desiderata it contributes: Parsimony, Plausibility, Coverage, Fidelity, Consistency, Continuity, and Efficiency. Since some metrics relate to more than one desideratum, we do not enforce an exclusive one-to-one mapping from metric to desiderata.

Contextuality: Second, we propose to distinguish metrics based on their evaluation context. We identify five different levels, each one pushing the contextuality of the evaluation further than the previous one. They are defined as follows:

I) Explanans-Centric: Evaluates only the explanans in relation to the raw input instance. It is entirely agnostic to the black-box model and does not rely on its predictions or behavior.
II) Model Observation: Assumes access to the black-box model's outputs or internal processes (e.g., activations), enabling evaluation based on observed behavior.
III) Input Intervention: Perturbs or modifies the input data, then observes the resulting predictions and explanantia to assess their changes.

IV) Model Intervention: Actively alters the black-box model itself, e.g., through retraining or parameter randomization.

V) A Priori Constrained: Requires specific setups of explananda, requiring particular data or models.

With each level, we further shift the evaluation focus from **In-Situ** to **Ex-Situ** evaluation[7]. The first three levels consist of In-Situ metrics, as they present an evaluation of concrete explanantia in the context of their original explanandum. The final two levels instead encompass Ex-Situ metrics, evaluating an explanation method more generally, either by imposing restrictions on the explanandum or by significantly altering the original context. This is similar to the common categorization of XAI methods into ante-hoc, in-hoc and post-hoc approaches (Carvalho et al., 2019; Zhang et al., 2021; Speith, 2022; Nauta et al., 2023; Bedi et al., 2024). Although Ex-Situ evaluations can serve to establish general comparisons between explanation methods, only In-Situ metrics provide information about the explanans in a given scenario.

Explanation Types: Third, we categorize VXAI metrics based on the accepted input. Apart from a few exceptions, all metrics are agnostic to the underlying black-box model or data format (e.g., tabular or image). Therefore, we do not consider this dimension separately. Instead, we draw inspiration from several XAI classification schemes, which propose a division based on the explanation approach or resulting explanans (Carvalho et al., 2019; Markus et al., 2021; Zhang et al., 2021; Speith, 2022; Nauta et al., 2023). We introduce and discuss the different explanation types in Subsection 5.1.2 below.

For XAI algorithms, a distinction exists between local and global methods (Zhang et al., 2021; Speith, 2022; Bedi et al., 2024), which we do not translate to VXAI metrics (unlike Robnik-Šikonja & Bohanec (2018)). Although not every metric is suited to calculate results for a single data input, the metrics can usually be adapted easily, e.g., calculating the change in logits for an individual input instead of calculating a change in accuracy over the dataset. Conversely, each metric that provides results for a single local explanans can be easily aggregated over a set of explanantia to obtain a global metric.

5.1.2 Explanation Types

We differentiate between the following explanation types: **Feature Attributions**, **Concepts**, **Examples**, **White-Box Surrogate Models**, and **Natural Language**. We semi-formally introduce each of the individual types. This choice also applies to our formulation of explanation types. Explanations are commonly categorized into their scope (Zhang et al., 2021; Speith, 2022; Bedi et al., 2024), where local explanations consider a single prediction and global explanations aim to elucidate the overall behavior of a black-box model. We include both local and global scope under the same explanation type.

Feature Attributions (FAs): A FA explanation returns a vector $e \in \mathbb{R}^d$, which usually (but not necessarily) has the same dimension as the input x. Each dimension represents an input feature, e.g., column in tabular data, (super-)pixel in images, or node in graphs. The value $e_j \in e$ then represents the relevance of the given feature towards the explained prediction. Depending on the underlying explanation algorithm, values can be positive or negative, and they may be inherently bounded to a given range or unbounded. Some FA methods assign continuous importance scores to each feature, while others produce binary or thresholded outputs that identify a subset of important features. Furthermore, Saliency Maps (which highlight important regions in the input, typically used in image-based tasks) present a special case of FAs, as features are not completely independent but exhibit spatial relationships. Usually, FAs are used as local explanations (Bach et al., 2015; Ribeiro et al., 2016; Lundberg, 2017; Shrikumar et al., 2017; Sundararajan et al., 2017), assigning feature importance for a single input prediction. However, they can also be global, e.g., showing the global impact of features (Lundberg, 2017; Molnar, 2020).

Concept Explanations (CEs): Similar to FAs, CEs can also be seen as a vector of importance scores. However, they are conceptually different as they represent a higher-level idea, above individual features, e.g., visual patterns rather than individual pixels. Therefore, there are usually fewer concepts than input features. Further, concepts are usually extracted from some intermediary representation or embedding inside the black-box model. While low-level features in FA are usually only meaningful for a given input (e.g., pixel),

[7] From latin for "in place" and "off-site", respectively.

concepts are meaningful on their own, and the same concept can be, and usually is, present in multiple inputs. Hence, CE strives a middle ground between local and global explanations, where the concepts themselves and their average contribution to prediction is a global explanation, while the detection of concepts in a single input is a local explanation (Kim et al., 2018).

Example Explanations (ExEs): ExEs are located in the input space, and therefore may consist of any type, including tabular data, images, graphs, and more. While the most prominent type of ExEs are Counterfactuals (minimally altered inputs that lead to a different prediction (Wachter et al., 2017)), other types such as Prototypes (typical examples representing a class (Kim et al., 2016)) or "Factuals" (altered instances that lead to the same prediction (Dhurandhar et al., 2019)) also belong to this category. ExEs may consist of a single explaining instance or can be a list of instances. Local ExE are common in the form of Counterfactuals, showing what would need to change to alter the individual prediction of a model to a desired outcome (Wachter et al., 2017; Karimi et al., 2020; Mothilal et al., 2020; Verma et al., 2024). Conversely, a global ExE could be a list of class prototypes or influential training samples (Kim et al., 2016; Koh & Liang, 2017; Molnar, 2020).

White-Box Surrogates (WBSs): WBSs aim to approximate the underlying black-box model. They achieve this through a surrogate model which is considered interpretable itself and therefore can serve as the explanans. This includes reconstructions of the full black-box model over the entire data space (Craven & Shavlik, 1995; Friedman & Popescu, 2008) but also local surrogates, which only serve as explanans for a given subset of instances (Ribeiro et al., 2016; 2018). Common types of inherently interpretable WBS are trees, rule-sets, or linear models.

Natural Language Explanations (NLEs): NLEs have been explored well before the rise of Large Language Models (LLMs), particularly through joint training setups that generate textual justifications alongside model predictions (Ras et al., 2022). However, the growing capabilities of LLMs have made NLEs increasingly prominent. Leveraging such models, the explanans may be generated alongside the prediction (Camburu et al., 2018; Wei et al., 2022) or post-hoc, potentially using a separate model (Bills et al., 2023). A more classical approach are template-based NLEs, where predefined building blocks are selected based on the results of other explanation methods (e.g., FAs) (Lucieri et al., 2022; Das et al., 2023). However, since these template-based approaches merely wrap an existing explanans in textual form, we do not consider them to be genuine NLEs. Instead, we propose to evaluate the underlying explanantia and explanations directly.

Notably, similar to the formulation of desiderata, our categorization scheme based on explanation types can be extended horizontally. We note that there is an overlap between different categories, as they represent both the final given explanans and the explanation process. LIME (Ribeiro et al., 2016) is a typical example, as its explanation fits a local WBS to generate a corresponding FA explanans. Similarly, there is a connection between CEs and other types, as WBSs can be leveraged to generate counterfactuals (Pornprasit et al., 2021), as can FAs (Ge et al., 2021; Albini et al., 2022). Rather than being a limitation, this overlap benefits our framework: it enables metrics designed for one explanation type to be applicable to others, facilitating broader reuse and comparison.

5.2 Identified Metrics

Now that we have established our classification scheme, we turn to the results of our literature review and report key statistics about the identified metrics.

We identified metrics from 362 individual sources. As some works proposed multiple variants, the total number of originally found metrics exceeded 400. These were grouped into 41 functionally distinct, aggregated metrics based on shared goals, methodology, or assumptions. Each was then categorized using the three-dimensional scheme introduced above. Since a single metric may serve multiple desiderata and apply to different explanation types, we organize them primarily by their contextuality level. Table 2 presents a complete overview of all metrics and their classification. A summary of key patterns follows below; for

Contextuality	Aggregated Metric	Desiderata							Explanation Type					#References
		Parsimony	Plausibility	Coverage	Fidelity	Continuity	Consistency	Efficiency	FA	ExE	CE	WBS	NLE	
I	(1) Explanans Size	✓							✓	✓	(✓)	✓	(✓)	51
	(2) Overlap	✓										✓		5
	(3) Explanans Cohesion	✓	✓						✓		(✓)			2
	(4) Minimality	✓	✓							✓				25
	(5) Autoencoder Plausibility		✓							✓				1
	(6) Diversity		✓							✓				6
	(7) Input Similarity		✓							✓				11
	(8) Input Contrastivity		✓		(✓)				✓	(✓)	(✓)	(✓)	(✓)	2
	(9) Actionability		✓		✓					✓				7
	(10) Model-Agnostic Explanation Consistency		✓				(✓)		✓	✓	(✓)	(✓)	(✓)	4
	(11) Input Coverage			✓					✓	✓	(✓)	✓	(✓)	8
	(12) Output Coverage			✓								✓		1
	(13) Output Mutual Information				✓				✓		✓			1
II	(14) Input Mutual Information	✓							✓		✓			1
	(15) Output Contrastivity		✓						✓	(✓)	(✓)	(✓)	(✓)	5
	(16) Output Similarity		✓							✓				1
	(17) Mutual Coherence		✓		(✓)				✓	(✓)	(✓)	(✓)	(✓)	19
	(18) Significance Check				✓				✓	(✓)	✓	(✓)	(✓)	15
	(19) (Counter-)Factual Relevance				✓					✓				1
	(20) Prediction Validity				✓					✓				18
	(21) Sufficency				✓						✓			2
	(22) Output Faithfulness				✓							✓		34
	(23) Internal Faithfulness				✓							✓		3
	(24) Setup Consistency						✓		✓	✓	(✓)	✓	(✓)	11
	(25) Hyperparameter Sensitivity						✓		✓	(✓)	(✓)	(✓)	(✓)	6
	(26) Execution Time							✓	✓	✓	(✓)	✓	(✓)	34
III	(27) Unguided Perturbation Fidelity				✓				✓					10
	(28) Guided Perturbation Fidelity				✓				✓		✓			75
	(29) Counterfactuability				✓							✓		1
	(30) Prediction Neighborhood Continuity				✓	✓						✓		1
	(31) Quantification of Unexplainable Features				(✓)	✓			✓		(✓)			2
	(32) Neighborhood Continuity					✓			✓	✓	(✓)	✓	✓	19
	(33) Adversarial Input Resilience					✓			✓	(✓)	(✓)	(✓)	(✓)	10
IV	(34) Model Parameter Randomization Test				✓				✓	(✓)	(✓)	(✓)	(✓)	5
	(35) Data Randomization Test				✓				✓	(✓)	(✓)	(✓)	(✓)	2
	(36) Retrained Model Evaluation				✓				✓					10
	(37) Influence Fidelity				✓					✓				2
	(38) Normalized Movement Rate					✓			✓					2
	(39) Adversarial Model Resilience					✓			✓	(✓)	(✓)	(✓)	(✓)	4
V	(40) GT Dataset Evaluation		✓		✓				✓		✓		✓	119
	(41) White Box Model Check				✓				✓	(✓)			(✓)	12

Table 2: The 41 aggregated metrics identified in our study, as presented in Appendix B. Each metric is classified according to its associated desiderata, applicable explanation types, and level of contextuality. ✓ indicates full alignment or reported usage in the literature, while (✓) denotes partial contribution or unreported but plausible applicability. The final column shows the number of references per metric.

detailed descriptions of each individual metric, see Appendix B. Four metrics deemed conceptually flawed or misaligned with any desideratum were excluded and are discussed in Appendix C.

Metric Popularity and References: On average, each metric is supported by 13.4 references (standard deviation 22.5), though the median is only 5. While a few metrics are backed by large reference sets (e.g., 119, 75, and 51 citations), eight metrics are supported by just one. This should not be interpreted as lack of relevance: our focus was on original proposals, not reuse or popularity. The most cited metric, Ground-Truth Dataset Evaluation, reflects the ubiquity of using annotated or synthetic datasets for evaluating explanation quality; an intuitive strategy with virtually unlimited implementation variants.

Desiderata: Most metrics target *Fidelity*, with 18 doing so directly and 3 more partially. *Plausibility* follows with 12 metrics. By contrast, *Efficiency*, *Coverage*, and *Consistency* are least represented, with only 1, 2, and 2 metrics respectively. The majority of metrics (32) address a single desideratum, while 4 target two equally and 5 contribute partially to a secondary one.

Explanation Types: A total of 25 metrics are applicable to FA methods. An equal number is available for ExE, although only 15 are directly supported by the literature; the remaining 10 are marked as potentially applicable (see (✓) in Table 2 and the opaque bars in Figure 5). For other explanation types, direct literature support is more limited: 11 metrics for WBS, 6 for CE, and just 2 for NLE. However, we consider many metrics adaptable even in the absence of published usage, increasing the totals to 22 for CE, 21 for WBS, and 17 for NLE. In total, 15 metrics are applicable (or adaptable) across all explanation types, and 19 are proprietary to a single type. Notably, no metric is exclusive to NLE, and only one is unique to CE. The largest overlap exists between FA and CE, with 21 metrics covering both. FA also has the highest number of adaptable metrics: while only 3 are exclusive to FA alone, 22 are shared with other types.

Contextuality: Most metrics fall into the less restrictive categories. Specifically, 13 are classified as *Contextuality I* (Explanans-Centric) and another 13 as *Contextuality II* (Model Observation), both of which do not require altering model or input. As contextual demands increase, metric availability declines: 7 fall into *Contextuality III* (Input Intervention), and 6 into *Contextuality IV* (Model Intervention). The most restrictive category, *Contextuality V* (A Priori Constrained), includes only two metrics that require specific setups with known ground-truth rationales, either via synthetic data or interpretable white-box models.

Desideratum-Contextuality Interactions: Metrics targeting the interpretability desiderata *Parsimony* and *Plausibility* are found almost exclusively in Contextualities I and II. Similarly, the technical desiderata *Coverage*, *Consistency*, and *Efficiency* are addressed only through observational strategies belonging to the same Contextualities of I and II. In contrast, all metrics measuring *Continuity* require intervention, either through modified inputs (Contextuality III) or altered model internals (Contextuality IV). *Fidelity*, in turn, is represented across all contextuality levels.

Contextuality-Explanation Interactions: There are no striking anomalies in how metrics available for specific explanation types are distributed across contextuality levels. The only exception is that no metric in Contextuality V targets WBS, likely because it is difficult to define a single ground-truth surrogate when multiple equivalent surrogates may exist.

Desideratum-Explanation Interactions: Every explanation type is covered by at least one metric for every desideratum, ensuring that all evaluation dimensions are, in principle, addressable regardless of the explanation form. The overall distribution across types is balanced. The only notable concentration occurs for *Plausibility* in ExE, with 10 applicable metrics. This may reflect the inherently interpretable nature of many ExE techniques (e.g., counterfactuals, prototypes), which naturally lend themselves to human-aligned plausibility assessments.

These patterns highlight the breadth of our framework. Overall, the proposed categorization encompasses a wide range of metrics across explanation types, desiderata, and evaluation contexts, offering a comprehensive foundation for structured VXAI assessment.

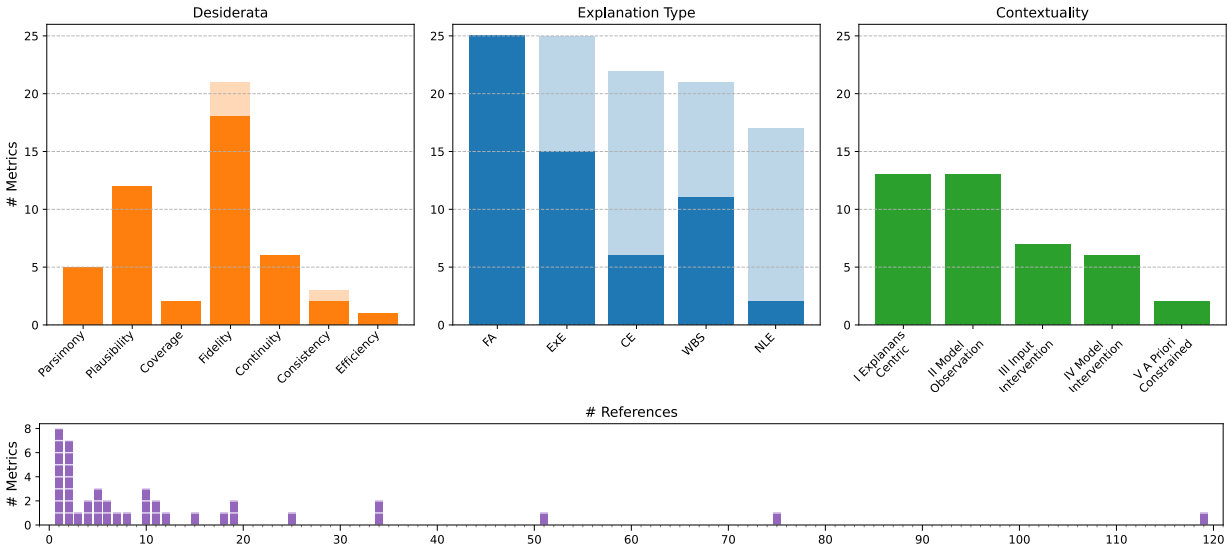


Figure 5: Overview of the distribution of metrics within the categorization scheme. Each metric may be associated with multiple desiderata and explanation types, but only a single level of contextuality. Light-colored bars indicate partial alignment with a desideratum. For explanation types, light bars denote cases where no usage has been reported in the literature, though the metric is considered adaptable. The bottom histogram shows the number of metrics grouped by their reference count.

6 Discussion

Finally, we reflect on trends observed in the identified VXAI metrics and offer considerations for their interpretation and future application.

General observations: Our framework reveals a wide and diverse landscape of VXAI metrics. While each aims to quantify a specific property of explanations, few metrics offer clear thresholds or benchmarks to determine what constitutes "good" explanation quality. This lack of consensus limits interpretability and comparability across studies. Furthermore, although several metrics address similar desiderata, explanation types, or contextuality levels, each typically serves a distinct niche. These differences often stem from variations in what aspect of a desideratum is targeted, or from adapting to specific explanation structures.

Emphasis on Fidelity: Fidelity stands out as the most frequently addressed desideratum, both in terms of metric count and methodological variety. This supports our perspective of Fidelity as a fundamental aspect of explanation quality (see Subsection 3.3). However, the range of proposed metrics also demonstrates that there is no universally accepted approach to quantifying it. The same holds for the desiderata Parsimony, Plausibility, and Continuity, each offering a range of possible evaluation metrics. In contrast, Coverage, Consistency, and Efficiency are addressed by fewer metrics. This is likely due to their more straightforward quantifiability; for instance, Coverage is trivially satisfied if explanantia exists, and Efficiency can be assessed via computation time. Portability, a part of Efficiency, is similarly important, but we identified no functionality-grounded metrics that are associated with it. Hence, it has to be viewed as a descriptive property of the method rather than something measurable on individual explanantia. More broadly, the number of metrics associated with a desideratum does not necessarily reflect how well it can be quantified. Some desiderata may be more reliably approximated by a single well-aligned metric than others by a diverse set of weaker proxies.

Explanation type focus: Among explanation types, FA is the most extensively studied, which is reflected both in literature prevalence and metric availability. This focus has also enabled the adoption of classical techniques from adjacent fields. For example, treating FA explanations as feature selectors allows the use

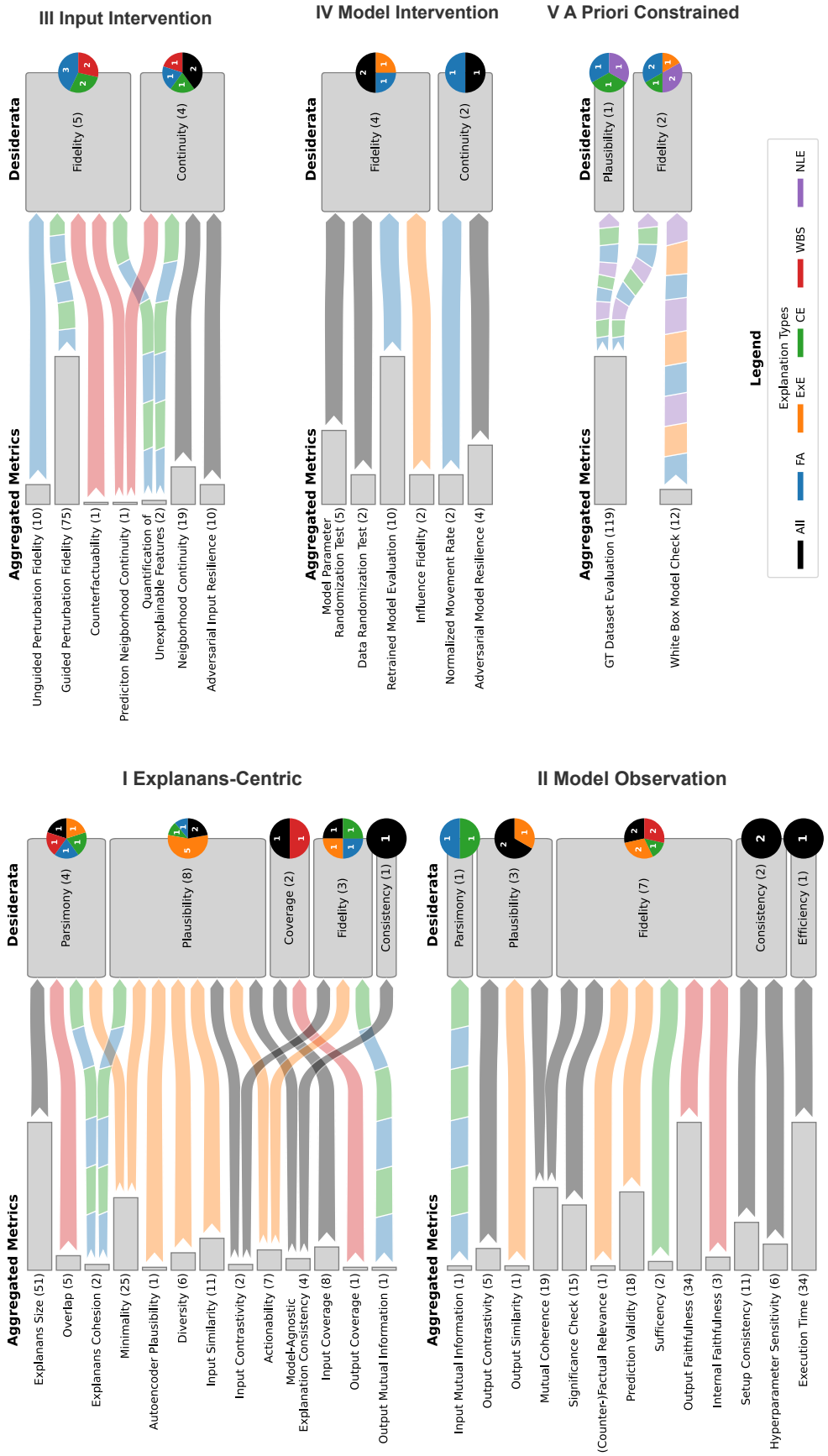


Figure 6: Overview of the 41 identified metrics, grouped by contextuality. Each metric is represented by a horizontal bar indicating the number of supporting references. Metrics are mapped to their associated desiderata, with arrow colors denoting the corresponding explanations type. These associations do not distinguish between full (✓) and partial (✓) alignment (unlike Table 2). For each desideratum, a pie chart summarizes the distribution of linked metrics by explanation type.

of established stability metrics from feature selection research, such as those used by Nogueira et al. (2018). Conversely, the explanation type NLE is still underrepresented in dedicated metrics. Nonetheless, our framework is flexible enough to incorporate emerging metrics in this and further areas.

Transferability across explanation types: A notable pattern is that many metrics are rarely applied beyond one or two explanation types. In our analysis, several were marked as "potentially applicable" to additional types, but lacked evidence of actual transfer. This suggests an isolated view of explanation types in current research. We view this as a missed opportunity, as generalizing and adapting metrics across different explanation types could foster a more integrated evaluation practice.

Metrics embedded in the explanation process: Some explanation types, notably ExE and WBS, commonly embed evaluation metrics as part of their generation objectives. For example, counterfactuals often optimize for plausibility and proximity directly during their construction (Wachter et al., 2017; Dandl et al., 2020; Kanamori et al., 2020). While this tight integration may enhance the generated explanans, it raises concerns about circularity in evaluation, especially in light of Goodhart's Law (Strathern, 1997): If a metric is optimized for, it may no longer serve as a reliable post-hoc assessment.

Subjectivity and the limits of automation: Metrics for Parsimony and Plausibility, while quantifiable, are ultimately shaped by subjective human interpretation. Functionality-grounded scores offer useful proxies that can be evaluated automatically and without costly human studies. However, they cannot replace human-centered assessments of whether an explanation is truly accessible or helpful in practice.

Caveats in specific metric designs: Several metric designs warrant closer scrutiny. Ground-truth-based metrics, specifically those relying on human-annotated rationales (see Metric 40), are widely used but have notable limitations. These include their tendency to reflect plausibility rather than explanation fidelity (Camburu et al., 2019; Carmichael & Scheirer, 2023), vulnerability to trivial baselines like central focus (Gu et al., 2019), and their inability to capture model reliance on context or background (Arras et al., 2022; Brandt et al., 2023). There's no reason to assume that the human rationale and the models' rationale are aligned (Khakzar et al., 2022), resulting in low scores for truthful explanations generated from an implausible model (Samek & Müller, 2019).

Likewise, metrics based on input perturbation or retraining (e.g., Metric 28, Metric 32, or Metric 36) can yield misleading results. Feature removal may not affect model output if redundant features are present (D'Amour et al., 2022), and retraining can alter the model in uncontrolled ways, obscuring meaningful evaluation (Hooker et al., 2019).

Metric aggregation: While most studies report multiple metric scores separately, some recent work proposes aggregating them into a single composite score (Farruque et al., 2021; Margot & Luta, 2021; Poppi et al., 2021; Bommer et al., 2024). Suggested methods include scaling by theoretical bounds, comparison to random or perfect baselines, and aggregation via weighted sums or harmonic means. Despite these proposals, the field lacks consensus on how to combine metrics meaningfully and robustly.

Evaluating the use of post-hoc explanations: These limitations prompt reflection on whether post-hoc explanation of black-box models is always justified. Some authors recommend comparing black-box performance to that of directly trained white-box models (Zhang et al., 2019b; Rosenfeld, 2021; Margot & Luta, 2021). If performance is comparable, choosing an interpretable model may be preferable, especially in high-stakes settings, as advocated by Rudin (2019). The same logic applies when considering surrogate models. They should be benchmarked against white-box models trained directly on the data (Barakat et al., 2010).

7 Conclusion, Limitations & Future Work

Conclusion

In this survey, we introduced a unified and comprehensive framework for the evaluation of XAI. We conducted a systematic review of the literature, identifying original metrics designed to assess explanation quality. These were grouped into 41 aggregated metrics and categorized using a three-dimensional scheme based on **desideratum**, **explanation type**, and **contextuality**.

Our analysis reveals a broad range of available metrics, covering diverse use cases and goals. At the same time, it highlights the lack of consensus regarding when and how specific metrics should be applied. Moreover, we find that many existing metrics are not yet adapted for all explanation types, indicating potential for extension and generalization.

Overall, our framework enables XAI practitioners to systematically analyze and compare explanations through an informed selection of suitable and comprehensive evaluation strategies. It also offers the possibility for extension, both by adding new metrics and by expanding the categorization scheme as the field evolves.

Limitations and Future Work

This work focuses exclusively on surveying and categorizing existing metrics. Although we provide structured comparisons and insights, we do not conduct empirical benchmarking. Future work should investigate how different metrics behave in practice, under which conditions they agree or contradict, and what trade-offs they impose. This would support the development of a more standardized benchmarking suite for explanation evaluation.

Establishing practical thresholds, studying alignment between metrics and desiderata, and exploring aggregation strategies remain open and important questions. Ideally, this could lead to a broadly accepted set of evaluation standards and protocols for verifying explanations across different models and tasks.

Lastly, while many metrics are labeled by us as potentially applicable to additional explanation types, their use remains unvalidated in literature. Future work should adapt and extend these metrics to underexplored explanation types, particularly for natural language and concept-based explanations.

Acknowledgment

A Queries

Scopus

```
TITLE-ABS-KEY (
  ("Explain* AI" OR "Explain* Artificial Intelligence" OR "Explain* ML" OR "
      Explain* Machine Learning" OR
  "Interpret* AI" OR "Interpret* Artificial Intelligence" OR "Interpret* ML"
     OR "Interpret* Machine Learning")
  AND
  ("Evaluation" OR "Metric" OR "Quantification")
  AND
  ("Survey" OR "Review")
)
```

IEEE Access

```
("All Metadata":"Explain* AI" OR "All Metadata":"Explain* Artificial
    Intelligence" OR "All Metadata":"Explain* ML" OR "All Metadata":"Explain*
    Machine Learning" OR
"All Metadata":"Interpret* AI" OR "All Metadata":"Interpret* Artificial
    Intelligence" OR "All Metadata":"Interpret* ML" OR "All Metadata":"
    Interpret* Machine Learning")
AND
("All Metadata":"Evaluation" OR "All Metadata":"Metric" OR "All Metadata":"
    Quantification")
AND
("All Metadata":"Survey" OR "All Metadata":"Review")
```

ACM

```
ANYWHERE:[
     ("Explain* AI" OR "Explain* Artificial Intelligence" OR "Explain* ML" OR "
         Explain* Machine Learning" OR
     "Interpret* AI" OR "Interpret* Artificial Intelligence" OR "Interpret* ML"
         OR "Interpret* Machine Learning")
     AND
     ("Evaluation" OR "Metric" OR "Quantification")
     AND
     ("Survey" OR "Review")
     ]
```

B VXAI Framework

B.1 Notation

The mathematical expressions used throughout the appendix are intended to clarify concepts. We do not enforce strict formalism as long as the notation remains unambiguous. We define the following symbols, which are used across all metrics:

The mathematical expressions used throughout the appendix are intended to clarify concepts. We do not enforce strict formalism as long as the notation remains unambiguous. In cases where clarity is not compromised, we slightly overload notation, but specific meanings are always disambiguated by context. While the notation is predominantly tailored to classification tasks, where the model assigns scores to discrete class labels, it can be adapted to suit regression settings or, in part, other paradigms. We define the following symbols, which are used across all metrics:

- Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input (data) and output (label) spaces, respectively.
- The black-box model is defined as a scoring function $\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that assigns a real-valued predictive score (e.g., logits) to each input–label pair.
- For a given input $x \in \mathcal{X}$, we write $\theta(x)$ as shorthand for the vector of scores across all labels in $\mathcal{Y}$, i.e., $\theta(x) := (\theta(x, y_1), \theta(x, y_2), \ldots, \theta(x, y_k))$.
- The predicted label is then given by $\hat{y}_x := \arg\max_{y \in \mathcal{Y}} \theta(x, y)$.
- We denote an arbitrary second input as $x' \in \mathcal{X}$ and a perturbed version of x as $\dot{x}$ (see Subsection B.2.1).
- Let $\mathcal{X}_y \subseteq \mathcal{X}$ be the set of all instances with ground-truth class y.

- An explanandum is a triple (θ, x, y), where θ is the model, $x \in \mathcal{X}$ is the input instance, and y is the label to be explained (which is often, but not necessarily, $\hat{y}$).
- An explanans for a given explanandum is denoted $e_{\theta,x,y}$, or simply e when unambiguous.
- For ExEs, we write $z := e^{(\text{ExE})}_{\theta,x,y^*_z}$, where $z \in \mathcal{X}$ is typically a counterfactual input such that $\hat{y}_z := \arg\max_{y \in \mathcal{Y}} \theta(z, y) = y^*_z$.
- For WBS, the surrogate model is denoted $\theta^e := e^{(\text{WBS})}_{\theta,x,\hat{y}}$. The surrogate returns a label prediction for arbitrary input via $\hat{y}^e_{x'} := \theta^e(x')$.

- $\delta(\cdot, \cdot)$ denotes an arbitrary distance or dissimilarity function, which may be applied to inputs, explanations, or predictions depending on context. Similarity measures (see Subsection B.2.3) may be applied by inverting them appropriately (e.g., via negation or reciprocal).
- $|\cdot|$ indicates size or cardinality; $\|\cdot\|_p$ denotes the L_p-norm.
- $\mathbb{1}[\cdot]$ is the indicator function, returning 1 if the condition holds and 0 otherwise.
- k refers to a generic parameter (e.g., number of selected features, neighbors, or examples); ϵ denotes a small threshold or tolerance constant.
- Φ_y denotes an autoencoder trained specifically on inputs from class y (i.e., $x \in \mathcal{X}_y$).

For convenience, we recall the abbreviations of the explanation types introduced in Subsection 5.1:

- Feature Attribution (FA)
- Concept Explanation (CE)
- Example Explanation (ExE)
- White-Box Surrogate (WBS)
- Natural Language Explanation (NLE)

Additionally, we adopt the following terminology: **Metrics** refer to the individual or aggregated VXAI criteria we evaluate and identified from the literature survey. **Measures** denote performance-scoring functions (e.g., accuracy, similarity, overlap), often also called metrics in literature; but we use the term "measure" to distinguish them clearly.

To improve readability while maintaining completeness throughout the Appendix, we move extensively long reference lists into footnotes using the notation:[a].

B.2 Helper Functions

The following components are recurring elements that serve as functional building blocks across multiple metrics. They can be understood as interchangeable parameters that shape how specific evaluation metrics are instantiated and interpreted.

B.2.1 Perturbation Approach

Perturbations are small changes typically applied to input features and are recurring components in metrics targeting Fidelity and Continuity. There exists a wide range of perturbation strategies, and the choice of approach can significantly affect both metric results and their interpretation (Brunke et al., 2020; Funke et al., 2022; Rong et al., 2022).

We distinguish first by the **Perturbation Scope**, i.e., the parts of the input that are modified. Perturbations can be applied at a fine-grained level (e.g., individual features) or on more structured, high-level groupings. For image data, scope definitions may involve aggregating pixels into fixed grids (Schulz et al., 2020) or segmenting into superpixels (Ribeiro et al., 2016; Kapishnikov et al., 2019; Rieger & Hansen, 2020). In time-series data, one may perturb fixed-length windows, with the target time-step at the beginning or middle (Schlegel et al., 2019; 2020). In topologically ordered domains (e.g., images, time-series, or graphs), adjacent features can be perturbed together, such as modifying the area surrounding a focal pixel (Samek et al., 2016; Brahimi et al., 2019). Higher-level approaches include perturbing Concepts (El Shawi et al., 2021) or internal activations associated with object parts (Zhang et al., 2019b).

Once the scope is defined, a variety of **Perturbation Functions** can be applied. In fact, any perturbation function might be suitable (Hameed et al., 2022; Schlegel & Keim, 2023). However, here we present some of the most common in literature. At the simplest level, features may be removed by setting them to zero or dropping them entirely, especially in structured domains such as graphs, text, or time-series data, as employed by many authors, (e.g., Bach et al. (2015); Ancona et al. (2017); Alvarez-Melis & Jaakkola (2018b); Chu et al. (2018); Arya et al. (2019); DeYoung et al. (2019); Schlegel et al. (2019); Cong et al. (2020); Singh et al. (2020); Warnecke et al. (2020); Bajaj et al. (2021); Faber et al. (2021); Singh et al. (2021); Jin et al. (2023)). Alternatively, features can be replaced by a fixed value, e.g., the per-channel or per-instance mean (Petsiuk, 2018; Schlegel et al., 2019; 2020; Hameed et al., 2022; Jin et al., 2023).

To generate less deterministic perturbations, authors propose adding random noise (e.g., Gaussian) or drawing values from uniform distributions (Yeh et al., 2019; Bhatt et al., 2020; Sturmfels et al., 2020; Bajaj et al., 2021; Funke et al., 2022; Veerappa et al., 2022). Other strategies leverage the spatial structure of the data: for instance, applying blurring or interpolation (Sturmfels et al., 2020; Rong et al., 2022), or reordering spatial regions (Schlegel et al., 2019; Chen et al., 2020). Instead of applying synthetic noise, values can be resampled from the marginal distribution of a feature, from its nearest neighbor, or even from an opposite-class example (Guo et al., 2018b; Hameed et al., 2022). Where influence regions are known, perturbations can be constrained to lie inside or outside these intervals (Velmurugan et al., 2021a).

In the NLP domain, word embeddings can be noised, or tokens substituted using synonym sets and domain knowledge (Yin et al., 2021). For image inputs, another strategy is cropping and resizing to emphasize or suppress local information (Dabkowski & Gal, 2017).

Finally, when perturbations are guided by a FA, their intensity can be scaled proportionally to the assigned importance scores (Chattopadhay et al., 2018; Guo et al., 2018b; Jung & Oh, 2021).

B.2.2 Normalization of Explanantia

Since FAs and CEs are typically represented as real-valued vectors, computed through various mechanisms, their value ranges are not inherently standardized. However, many metrics either explicitly require the

[a] Author et al.

explanans to lie within a fixed range or implicitly assume comparability across explanantia, making normalization a necessary preprocessing step. A widely used normalization method is Min-Max Scaling (Binder et al., 2023; Brandt et al., 2023), which maps all values into a fixed interval (typically $[0, 1]$). Alternative strategies include normalization based on the square root of the average second-moment estimate, offering robustness to outliers and variance shifts (Binder et al., 2023). This limited selection can be extended through any suitable normalization approach.

B.2.3 Similarity Measures

Across various metrics, it is necessary to calculate similarities or distances between two explanantia, especially when concerned with the desideratum of Continuity and in metrics relying on Ground-Truth evaluations. Throughout the reported literature, various approaches have been reported, which differ based on the type of explanation. While some measures directly compute similarity, others quantify distance or disparity. In this work, we adopt a similarity-based framing, either directly or by transforming distance measures, to ensure that higher values uniformly indicate greater explanatory agreement. Analogously, we can compute the similarity between explananda, adopting measures that are presented in the following.

For **FA**s, similarity may be measured using arbitrary inverted distance or loss measures (e.g., L_p, MSE, cosine distance, JS-Divergence, or Bhattacharyya Coefficient), potentially normalized (e.g., by standard deviation)[b]. Alternatively, rank correlation measures such as Spearman's or Kendall's Tau can be applied[c]. When binarizing FA outputs through thresholding, feature-wise evaluation measures such as accuracy, precision, F_1, or AUROC are commonly used[d]. Similarly, IoU can be calculated over binarized features[e], or top-k intersection measures may be used[f]. For saliency maps, specialized similarity measures are available, such as SSIM[g], Earth Movers Distance (Park et al., 2018; Wu & Mooney, 2018), Normalized Cross Correlation (Baumgartner et al., 2018; Bass et al., 2020), or Mutual Information (Sun et al., 2023). While primarily established for FAs, similar similarity functions can be naturally applied to **CE**-based explanations as well.

For **WBS**s and **ExE**s, the choice of similarity measure depends strongly on the underlying model or domain. For linear predictive models, coefficient mismatch is a common choice (Lakkaraju et al., 2020), whereas rule- and tree-based explanantia may be compared by their rule overlap, feature usage, or node structures[h].

NLEs can be compared using standard natural language processing measures[i]. Those include for instance BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), or SPICE (Anderson et al., 2016). In addition, several measures have been proposed specifically for evaluating natural language explanations (Xie et al., 2021; Du et al., 2022; Rodis et al., 2024; Park et al., 2018).

When comparing similarities over multiple instances, the most natural aggregation is to compute the mean similarity (Fan et al., 2020; Fouladgar et al., 2022; Yeh et al., 2019). Depending on the evaluation goal, alternative aggregation strategies may offer more informative insights. For example, worst-case stability, defined as the minimum similarity across inputs, can be used to quantify robustness[j].

[b] Alvarez-Melis & Jaakkola (2018a;b); Chu et al. (2018); Wu & Mooney (2018); Jain & Wallace (2019); Jia et al. (2019); Mitsuhara et al. (2019); Pope et al. (2019); Trokielewicz et al. (2019); Yeh et al. (2019); Zhang et al. (2019a); Jia et al. (2020); Agarwal et al. (2022a); Atanasova et al. (2022); Dai et al. (2022); Fouladgar et al. (2022); Agarwal et al. (2023); Huang et al. (2023a); Nematzadeh et al. (2023)

[c] Das et al. (2017); Adebayo et al. (2018); Chen et al. (2019b); Dombrowski et al. (2019); Ghorbani et al. (2019); Nguyen & Martínez (2020); Rajapaksha et al. (2020); Sanchez-Lengeling et al. (2020); Liu et al. (2021a); Yin et al. (2021); Krishna et al. (2022); Huang et al. (2023a)

[d] Chen et al. (2018a); Yang et al. (2018b); Jia et al. (2019; 2020); Sanchez-Lengeling et al. (2020); Bykov et al. (2021); Joshi et al. (2021); Park & Wallraven (2021); Amoukou et al. (2022); Chen et al. (2022); Funke et al. (2022); Tjoa & Guan (2022); Wilming et al. (2022); Agarwal et al. (2023)

[e] Oramas et al. (2017); Fan et al. (2020); Kim et al. (2021); Situ et al. (2021); Vermeire et al. (2022)

[f] Ghorbani et al. (2019); Mishra et al. (2020); Rajapaksha et al. (2020); Warnecke et al. (2020); Amparore et al. (2021); Bajaj et al. (2021)

[g] Adebayo et al. (2018); Dombrowski et al. (2019); Rebuffi et al. (2020); Graziani et al. (2021); Sun et al. (2023)

[h] Bastani et al. (2017); Guidotti et al. (2019); Lakkaraju et al. (2020); Rajapaksha et al. (2020); Margot & Luta (2021)

[i] Camburu et al. (2018); Chuang et al. (2018); Liu et al. (2018a); Wu & Mooney (2018); Chen et al. (2019d); Rajani et al. (2019); Wickramanayake et al. (2019); Li et al. (2020a); Sun et al. (2020); Jang & Lukasiewicz (2021); Atanasova (2024b)

[j] Alvarez-Melis & Jaakkola (2018a;b); Yeh et al. (2019); Yin et al. (2021); Fouladgar et al. (2022)

B.3 Metrics

B.3.I Explanans-Centric

The following metrics directly evaluate the explanans, potentially looking into the data inputs but without having any access to the underlying model.

Metric I.1: Explanans Size

Desiderata: Parsimony
Explanation Type: FA, ExE, WBS, (CE, NLE) **References:** (51)
Craven & Shavlik (1995); Stefanowski & Vanderpooten (2001); Nauck (2003); Alonso et al. (2008); Augasta & Kathirvalavakumar (2012); Lakkaraju et al. (2016); Samek et al. (2016); Zilke et al. (2016); Lakkaraju et al. (2017); Guidotti et al. (2018); Hara & Hayashi (2018); Rustamov & Klosowski (2018); Wang et al. (2018b); Wang (2018); Wang et al. (2018a); Wu et al. (2018a;b); Deng (2019); Evans et al. (2019); Fong et al. (2019); Guidotti et al. (2019); Ignatiev et al. (2019); Lakkaraju et al. (2019); Polato & Aiolli (2019); Pope et al. (2019); Shakerin & Gupta (2019); Slack et al. (2019); Topin & Veloso (2019); Verma & Ganguly (2019); Wang (2019); Yoo & Sael (2019); Bhatt et al. (2020); Chalasani et al. (2020); Molnar et al. (2020); Nguyen & Martínez (2020); Panigutti et al. (2020); Rajapaksha et al. (2020); Rawal & Lakkaraju (2020); Stepin et al. (2020); Warnecke et al. (2020); Wu et al. (2020); Liu et al. (2021b); Margot & Luta (2021); Moradi & Samwald (2021); Poppi et al. (2021); Rosenfeld (2021); Samek et al. (2021); Dai et al. (2022); Funke et al. (2022); Huang et al. (2023b); Stevens & De Smedt (2024)

The size of an explanans $|e|$ is a common indicator of its complexity. Smaller or more compact explanantia are generally easier to understand and more plausible to human users. The exact method to measure size depends on the explanation type and context.
A generally applicable method is to compute the file size (in bytes) of the explanans, based on the assumption that sparse explanantia can be more easily compressed (Samek et al., 2016; 2021).
For **WBS**, size is typically measured via structural properties of the surrogate model: tree-based models are assessed by depth, number of nodes, or number of leaves[a], while rule-based systems are evaluated using the number of rules or predicates per rule[b]. For explanation-graphs, the number of nodes and edges can serve as a proxy for size (Rustamov & Klosowski, 2018; Topin & Veloso, 2019). Conversely, the (relative) number of instances covered per rule can also express parsimony (Deng, 2019; Guidotti et al., 2019).
In **FA**, size is typically based on the number of relevant features. This may be computed via:

- The L_0 norm[c],
- A count of features exceeding a relevance threshold[d],
- Normalization by input dimensionality, e.g., in graph settings (Pope et al., 2019),
- Integration over multiple thresholds to form a size curve (Warnecke et al., 2020), or
- Threshold-free statistics like entropy (Samek et al., 2016; Bhatt et al., 2020; Funke et al., 2022) and Gini index (Chalasani et al., 2020).

While not common in the literature, for **CE**, similar counting mechanisms may be applied. One may count concepts exceeding a relevance threshold or use the total number of tested concepts (which is not tied to input size).
In **ExE**, size naturally corresponds to the number of examples used in the explanans (Nguyen & Martínez, 2020; Huang et al., 2023b).
For **NLE**, the number of words or sentences provides a straightforward measure of size.
In general, size can be measured per-instance or aggregated over the dataset. For instance, one can report the number of predicates in the rule explaining a single instance (Augasta & Kathirvalavakumar, 2012), or aggregate the number of predicates across a global rule set with statistics such as the average (Lakkaraju et al., 2016; 2017; 2019), sum (Margot & Luta, 2021; Moradi & Samwald, 2021), or maximum (Moradi & Samwald, 2021). Similarly, rule counts can be aggregated over classes (e.g., average number per class (Nauck, 2003)). Optional adjustments include applying a tolerance margin

(e.g., $\max(0, |e| - k)$ (Rosenfeld, 2021)) or aggregating over multiple class-specific explanantia per instance (Pope et al., 2019).

[a]Craven & Shavlik (1995); Alonso et al. (2008); Guidotti et al. (2018); Hara & Hayashi (2018); Wu et al. (2018b); Evans et al. (2019); Slack et al. (2019); Yoo & Sael (2019); Molnar et al. (2020); Rawal & Lakkaraju (2020); Wu et al. (2020)

[b]Craven & Shavlik (1995); Stefanowski & Vanderpooten (2001); Nauck (2003); Alonso et al. (2008); Augasta & Kathirvalavakumar (2012); Lakkaraju et al. (2016); Zilke et al. (2016); Lakkaraju et al. (2017); Wang (2018); Wu et al. (2018a); Deng (2019); Lakkaraju et al. (2019); Polato & Aiolli (2019); Shakerin & Gupta (2019); Wang (2019); Panigutti et al. (2020); Rajapaksha et al. (2020); Stepin et al. (2020); Margot & Luta (2021); Moradi & Samwald (2021); Rosenfeld (2021)

[c]Wang et al. (2018a;b); Fong et al. (2019); Polato & Aiolli (2019); Verma & Ganguly (2019); Nguyen & Martínez (2020); Poppi et al. (2021); Rosenfeld (2021); Stevens & De Smedt (2024)

[d]Ignatiev et al. (2019); Pope et al. (2019); Warnecke et al. (2020); Liu et al. (2021b); Dai et al. (2022)

Metric I.2: Overlap

Desiderata: Parsimony
Explanation Type: WBS **References:** (5)
Lakkaraju et al. (2016; 2017; 2019); Moradi & Samwald (2021); Hosain et al. (2024)

The overlap within a rule-based explanans (e.g., rule sets or decision trees) can be measured with respect to either the input space or the rules themselves. A lower degree of overlap is generally preferred, as it implies more distinct, non-redundant rules and enhances interpretability.

- **Input Overlap**: Measures how often input instances are covered by multiple rules or decision paths (Lakkaraju et al., 2016; 2017; 2019; Hosain et al., 2024). High overlap may indicate redundant or conflicting logic. Overlap can also be broken down by class label to distinguish intra-class from inter-class coverage.
- **Rule Overlap**: Measures how many rules share identical or highly similar predicates (Moradi & Samwald, 2021), indicating structural redundancy within the rule set.

Metric I.3: Explanans Cohesion

Desiderata: Parsimony, Plausibility
Explanation Type: FA,(CE) **References:** (2)
Fong & Vedaldi (2017); Saifullah et al. (2024)

In domains where input features are ordered (e.g., images or time series, as opposed to tabular data), we expect relevant information to be located in a smooth and coherent region of the input. This property can be evaluated in two complementary ways.
First, Fong & Vedaldi (2017) assess the locality of the explanans: a higher degree of cohesion is achieved when relevant information lies within a small, contiguous region of interest, making the explanans more condensed and easier to interpret. In saliency maps, this can be quantified by computing the area of the smallest bounding box that encloses the thresholded explanans.
Second, Saifullah et al. (2024) assess the smoothness of the explanans: a more continuous attribution map is often easier to understand. This can be measured by summing the absolute differences in attribution between neighboring features (e.g., in x- and y-directions for images). A higher score indicates a more fragmented, less interpretable explanans.
Both variants are applicable to data with ordered features and extend naturally to temporal domains. While originally proposed for FAs, these metrics can also be applied to CE when concept-based saliency maps are available (e.g., as presented by Lucieri et al. (2020)).

Metric I.4: Minimality

Desiderata: Parsimony, Plausibility
Explanation Type: ExE **References:** (25)
Tolomei et al. (2017); Wachter et al. (2017); Karlsson et al. (2018); Zhang et al. (2018c); Guidotti et al. (2019); Albini et al. (2020); Artelt & Hammer (2020); Dandl et al. (2020); Kanamori et al. (2020); Karimi et al. (2020); Le et al. (2020); Mothilal et al. (2020); Pawelczyk et al. (2020); Ramon et al. (2020); Sharma et al. (2020); Abrate & Bonchi (2021); Hvilshøj et al. (2021); Pawelczyk et al. (2021); Rasouli & Yu (2021); Van Looveren & Klaise (2021); Albini et al. (2022); Chou et al. (2022); Bayrak & Bach (2023a); Huang et al. (2023b); Verma et al. (2024)

To ensure that counterfactuals are understandable and believable, the induced changes should be minimal. This can be assessed by evaluating both the proximity of the counterfactual to the original instance and the sparsity of the changes.
Proximity is commonly measured as the distance $\delta(x, z)$ between the counterfactual z and the original instance x. A variety of distance measures are used, with the choice significantly impacting the results (Wachter et al., 2017; Artelt & Hammer, 2020; Bayrak & Bach, 2023a; Verma et al., 2024) Common options include (feature-wise weighted) L_p distances (especially L_2)[a], as well as Jaccard or cosine distance (Tolomei et al., 2017), or combinations thereof (Karimi et al., 2020; Hvilshøj et al., 2021). Other notable choices are Mahalanobis distance (Artelt & Hammer, 2020; Kanamori et al., 2020; Verma et al., 2024), Gower distance (Dandl et al., 2020; Karimi et al., 2020), or feature-wise cumulative density-based distances (Pawelczyk et al., 2020). Dataset-specific alternatives include computing the quantile-shift per feature (Albini et al., 2022). Different data types require suitable distance measures, e.g.: for graphs, the symmetric difference of adjacency matrices or cosine similarity between node features (Abrate & Bonchi, 2021); for images, the inverse SSIM score (Sharma et al., 2020); and for time series, distances measured per time step (Karlsson et al., 2018).
Sparsity, in contrast, concerns the number of features changed rather than the extent of change. It is often quantified via the L_0 norm, that is, the number (or fraction) of altered features[b].
Bridging sparsity and proximity, measures like L_1 (Wachter et al., 2017) or elastic-net regularization (Van Looveren & Klaise, 2021) are sometimes used. Depending on the representation, alternative definitions may apply, for example, counting the number of changed rules in rule-based counterfactuals (Guidotti et al., 2019).

[a] Wachter et al. (2017); Artelt & Hammer (2020); Le et al. (2020); Mothilal et al. (2020); Sharma et al. (2020); Rasouli & Yu (2021); Chou et al. (2022); Verma et al. (2024)
[b] Albini et al. (2020); Dandl et al. (2020); Le et al. (2020); Mothilal et al. (2020); Ramon et al. (2020); Pawelczyk et al. (2021); Bayrak & Bach (2023a); Verma et al. (2024)

Metric I.5: Autoencoder Plausibility

Desiderata: Plausibility
Explanation Type: ExE **References:** (1)
Van Looveren & Klaise (2021)

Autoencoders are trained to capture the underlying structure of the dataset. Based on this idea, Van Looveren & Klaise (2021) propose evaluating the plausibility of counterfactual examples via reconstruction errors from class-specific and general-purpose autoencoders.
Two specific scores are introduced:

- **IM1** (Class-specific reconstruction comparison): This metric compares how well a counterfactual z is reconstructed by an autoencoder trained on the target class ($\Phi_{y_z^*}$) versus the originally predicted class ($\Phi_{\hat{y}_x}$): $\frac{\|z-\Phi_{y_z^*}(z)\|_2^2}{\|z-\Phi_{\hat{y}_x}(z)\|_2^2+\epsilon}$.
  A low IM1 score implies that the counterfactual is more representative of the target class than of its original class, indicating class-specific plausibility.

- **IM2** (General manifold plausibility): This score compares the reconstructions from a general autoencoder $\Phi_{\mathcal{Y}}$ and a class-specific one: $\frac{\|\Phi_{\mathcal{Y}}(z)-\Phi_{y_z^*}(z)\|_2^2}{\|z\|_1+\epsilon}$.
  A low IM2 score implies that the counterfactual aligns well with both the general data manifold and the target class distribution, thus indicating higher plausibility.

While these metrics provide valuable insight into how realistic or semantically valid counterfactuals are, they come with caveats. Training multiple autoencoders (e.g., per class) introduces significant computational overhead. Further, reconstruction quality may vary across classes, and Hvilshøj et al. (2021) showed that autoencoders can be sensitive to small perturbations, potentially undermining the consistency of the plausibility assessment.

Metric I.6: Diversity

Desiderata: Plausibility
Explanation Type: ExE **References:** (6)
Wachter et al. (2017); Karimi et al. (2020); Mothilal et al. (2020); Nguyen & Martínez (2020); Stepin et al. (2020); Verma et al. (2024)

In cases where multiple counterfactuals are generated for a single instance, they should be diverse to offer distinct and meaningful alternatives (Wachter et al., 2017; Stepin et al., 2020; Verma et al., 2024).
Diversity is typically quantified by computing pairwise distances among the set of counterfactuals, either as the average pairwise distance (Mothilal et al., 2020; Nguyen & Martínez, 2020), or as the number of counterfactual pairs that exceed a predefined distance threshold (Karimi et al., 2020).

Metric I.7: Input Similarity

Desiderata: Plausibility
Explanation Type: ExE **References:** (11)
Laugel et al. (2019a;b); Mahajan et al. (2019); Singla et al. (2019); Artelt & Hammer (2020); Dandl et al. (2020); Kanamori et al. (2020); Delaney et al. (2021); Pawelczyk et al. (2021); Rasouli & Yu (2021); Smyth & Keane (2022)

To ensure that counterfactuals remain realistic and trustworthy, they should lie close to the true data manifold. Counterfactuals that deviate significantly from the distribution of training data are unlikely to be plausible. Two principal approaches are used to evaluate this alignment: direct estimation of data conformity and distance-based proximity to the training data.
In the **direct approach**, several statistical or unsupervised methods determine whether the counterfactual is an outlier. This includes calculating its likelihood under a kernel density estimator (Artelt & Hammer, 2020) or from known data distributions in synthetic setups (Mahajan et al., 2019). Other methods include Local Outlier Factor (Kanamori et al., 2020; Delaney et al., 2021) and Isolation Forests (Delaney et al., 2021), which can be applied either across the full dataset or restricted to samples of the counterfactual's target class (Artelt & Hammer, 2020). Pawelczyk et al. (2021) assess how many of a counterfactual's nearest neighbors share its class label.
Alternatively, proximity can be evaluated through **distance-based** approaches, relying on any suitable distance measure (or inverted similarity function, see Subsection B.2.3). These include distances to the k nearest neighbors in the training data (Laugel et al., 2019a; Dandl et al., 2020), or to the nearest training instances that share the same class label as the counterfactual or the original instance (Rasouli & Yu, 2021; Smyth & Keane, 2022). Specific domains may require tailored metrics such as the Fréchet Inception Distance (FID) for images (Singla et al., 2019).
Raw distance scores are often used directly (Dandl et al., 2020), but some methods normalize distances, for example, by comparing the counterfactual–input distance $\delta(x, z)$ to average distances

between random pairs $\delta(x', x'')$ in the dataset (Laugel et al., 2019a;b), or to the original instance–counterfactual distance (Smyth & Keane, 2022).

Metric I.8: Input Contrastivity

Desiderata: Plausibility, (Fidelity)
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (2)
Honegger (2018); Pornprasit et al. (2021)

Explanantia should be specific to their corresponding explananda and not overly generic. That is, distinct inputs should typically yield distinct explanantia.
To assess this, we calculate the fraction of instances or instance pairs that result in different explanantia (Honegger, 2018; Pornprasit et al., 2021). A higher fraction of distinguishable explanantia indicates greater contrastivity and, by extension, higher plausibility.
While this metric has been primarily proposed for FAs, it may be applicable to other explanation types as well.

Metric I.9: Actionability

Desiderata: Plausibility, Fidelity
Explanation Type: ExE **References:** (7)
Mahajan et al. (2019); Karimi et al. (2020); Le et al. (2020); Pawelczyk et al. (2021); Ma et al. (2022); Smyth & Keane (2022); Verma et al. (2024)

For a counterfactual to be helpful, it should only introduce changes that are feasible or allowed. This is often enforced by defining explicit constraints, such as:

- Immutable features (e.g., sex, age) (Karimi et al., 2020; Pawelczyk et al., 2021; Smyth & Keane, 2022; Verma et al., 2024), or
- Value boundaries for individual features (Karimi et al., 2020; Le et al., 2020; Ma et al., 2022; Smyth & Keane, 2022).

The actionability of a counterfactual set can then be assessed by calculating the fraction of counterfactuals that satisfy all constraints[a], or by computing the harmonic mean of satisfied constraints per counterfactual (Mahajan et al., 2019).

[a] Karimi et al. (2020); Le et al. (2020); Pawelczyk et al. (2021); Ma et al. (2022); Smyth & Keane (2022)

Metric I.10: Model-Agnostic Explanation Consistency

Desiderata: Plausibility, (Consitency)
Explanation Type: FA, ExE, (CE, WBS, NLE) **References:** (4)
Fan et al. (2020); Nguyen et al. (2020); Hvilshøj et al. (2021); Jiang et al. (2023)

To assess whether explanations reflect generalizable patterns rather than model-specific artifacts (such as adversarial shortcuts in counterfactual explanations), several authors evaluate explanations across different models trained on the same dataset.
In general, the approach involves training additional black-box models (oracles) on the same data and task. These oracles are then used to evaluate explanations from the original model.
One strategy is to directly compute the similarity between explanantia generated by different models for the same input (Fan et al., 2020). While this has been reported for FA, it is applicable to any explanation type, provided a suitable similarity metric is chosen (see Subsection B.2.3).
Alternatively, explanations from the original model are evaluated using the oracles:

- Nguyen et al. (2020) use perturbation-based evaluation (e.g., Metric 28) on a secondary model to assess whether the explanation highlights features that are generally important across models.
- Hvilshøj et al. (2021) propose that counterfactuals should change the prediction in both the original model and the oracle. Only such counterfactuals are considered plausible.
- An extension of this approach trains one oracle per class and computes the Jensen-Shannon-Divergence between its predictions on the original and counterfactual input. Ideally, only the target and original class should exhibit strong divergence (Hvilshøj et al., 2021).
- Jiang et al. (2023) define a neighborhood of models via small weight perturbations and count how many counterfactuals remain valid across all neighbors.

This approach might be extended to other explanation types. However, its interpretive strength remains limited: it is unclear whether explanations should be similar across models, as differing model architectures may learn distinct (yet valid) rationales. As a result, high or low agreement does not always reflect explanation quality, making this metric inherently context-dependent.

Metric I.11: Input Coverage

Desiderata: Coverage
Explanation Type: FA, ExE, WBS, (CE,NLE **References:** (8)
Lakkaraju et al. (2016; 2017; 2019); Ribeiro et al. (2018); Rawal & Lakkaraju (2020); Warnecke et al. (2020); Moradi & Samwald (2021); Huang et al. (2023b)

An explanation method should ideally be capable of generating a valid explanans for every input instance, regardless of its position on the data manifold. To assess this, we calculate the fraction of inputs for which the method provides a valid result.
The notion of "valid result" is intentionally broad and depends on the explanation type and the use case. A few prominent definitions include:

- For perturbation-based FAs like LIME (see Ribeiro et al. (2016)), a valid explanans can be given if a sufficient number of perturbed samples belong to the target or opposite class, enabling a reliable local approximation (Warnecke et al., 2020).
- For rule-based or tree-based WBSs, coverage is the number of instances that are captured by at least one rule or decision path (Lakkaraju et al., 2016; 2017; 2019; Ribeiro et al., 2018; Rawal & Lakkaraju, 2020), optionally restricted to rules associated with the correct class (Moradi & Samwald, 2021).
- For global ExEs, coverage may be defined as the number of inputs that have at least one sufficiently similar explaining instance within a pre-defined distance (Huang et al., 2023b).

Metric I.12: Output Coverage

Desiderata: Coverage
Explanation Type: WBS **References:** (1)
Lakkaraju et al. (2016)

Inherently interpretable WBS models should be capable of producing predictions for all possible output classes. Lakkaraju et al. (2016) evaluate the model's class-level coverage by calculating the fraction of target classes that appear in the explanans. This could, for instance, be the number of classes that are assigned to at least one rule in a rule set or to a leaf in a decision tree.

Metric I.13: Output Mutual Information

Desiderata: Fidelity
Explanation Type: FA,CE **References:** (1)
Nguyen & Martínez (2020)

When high-level features (such as concepts or aggregated input features) are provided by an explanans, they should capture as much information as possible about the model's output. In turn, the prediction should be reflected in the explanans, implying a high degree of mutual dependency between both.
To quantify this relationship, Nguyen & Martínez (2020) leverage the *Mutual Information (MI)* (see Cover (1999)) between the explanans and the output: $\mathbf{MI}(e_{\theta,x,\hat{y}}, \theta(x))$. Since MI is symmetric, a high score indicates that the explanation both reflects (i.e. correctness) and encompasses (i.e., completeness) the relevant reasoning of the model.

B.3.II Model Observation

This set of metrics requires access to the model to obtain information about the model's output or its internal activations.

Metric II.14: Input Mutual Information

Desiderata: Parsimony
Explanation Type: FA,CE **References:** (1)
Nguyen & Martínez (2020)

When high-level features are provided by an explanans, they should ideally represent an abstracted, human-understandable form of the input, omitting unnecessary detail. This promotes parsimony by simplifying the input space. Such features can be obtained either through feature selection in FAs or through concept-level representations in CEs.
To quantify this abstraction, Nguyen & Martínez (2020) propose measuring the *Mutual Information (MI)* (see (Cover, 1999)) between the input and the explanans: $\mathbf{MI}(x, e_{\theta,x,\hat{y}})$ A lower mutual information score suggests that the explanans captures less granular input detail and thus constitutes a more compact and parsimonious representation.

Metric II.15: Output Contrastivity

Desiderata: Plausibility
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (5)
Nie et al. (2018); Pope et al. (2019); Li et al. (2020c); Rebuffi et al. (2020); Sixt et al. (2020)

To enhance plausibility, explanations should be class-discriminative, that is, they should differ depending on the target class. In image classification, for instance, the explanantia supporting a "car" label should highlight different regions than those supporting "dog". This makes it easier for humans to understand the specific rationale for each class.
Class discriminativeness is commonly measured by comparing explanantia generated for different classes on the same input. Several comparison setups have been proposed: between the most and least likely classes (Li et al., 2020c; Rebuffi et al., 2020), the top predicted vs. a randomly chosen class (Sixt et al., 2020), or simply between the two classes in binary settings (Pope et al., 2019).
Any suitable similarity or distance measure can be used to assess the degree of overlap between explanantia, such as the L_0 or L_2 distance (Nie et al., 2018; Pope et al., 2019), rank correlation (Rebuffi et al., 2020), or SSIM for image domain (Sixt et al., 2020). A lower similarity implies a clearer distinction between the rationales for each class.

Although the reported implementations target FA, the idea of class-discriminative explanations can be naturally extended to other type, wherever explanations can be generated for multiple class hypotheses.

Metric II.16: Output Similarity

Desiderata: Plausibility
Explanation Type: ExE **References:** (1)
Plumb et al. (2020)

To ensure plausibility, counterfactuals should yield output distribution similar to real instances of the target class. Plumb et al. (2020) evaluate whether each counterfactual z matches the output activation of at least one training sample, i.e.: $\exists x' \in \mathcal{X}_{y_z^*} \ : \ \delta\big(\theta(x'), \theta(z)\big) < \epsilon$
This confirms that the counterfactual aligns with typical model behavior for that class.

Metric II.17: Mutual Coherence

Desiderata: Plausibility, (Fidelity)
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (19)
Selvaraju et al. (2017); Guo et al. (2018a); Ancona et al. (2019); Fernando et al. (2019); Fusco et al. (2019); Jain & Wallace (2019); Zhang et al. (2019a); Marques-Silva et al. (2020); Nguyen & Martínez (2020); Wang et al. (2020b); Warnecke et al. (2020); Graziani et al. (2021); Malik et al. (2021); Rajbahadur et al. (2021); Krishna et al. (2022); Mercier et al. (2022); Jin et al. (2023); Duan et al. (2024); Tekkesinoglu & Pudas (2024)

Several authors propose to evaluate their explanations by comparing them against other XAI methods that are assumed to produce trustworthy or well-understood outputs. Although only reported for FAs, this approach is applicable to any explanation type, as long as a suitable reference XAI method is available and an appropriate similarity metric can be defined.
A common reference is the Shapley value framework (see Shapley (1953)), often operationalized via SHAP (see Lundberg (2017)) due to its solid theoretical grounding. FAs are frequently compared to SHAP estimates using similarity measures[a]. In the context of saliency maps, Image Occlusion (see Zeiler & Fergus (2013)) is similarly treated as a proxy ground truth (Selvaraju et al., 2017). Comparisons against further explanation methods are also reported in literature[b].
Other works compute similarity across multiple explanation methods, assuming that high agreement between methods reflects convergence toward a reliable explanans[c]. Any suitable similarity measure may be applied (see Subsection B.2.3 for an overview).
However, this evaluation approach is not without criticism. Kumar et al. (2020) caution that validating one explanation method using another may propagate shared biases or assumptions, providing limited evidence for actual correctness.

[a] Ancona et al. (2019); Zhang et al. (2019a); Malik et al. (2021); Jin et al. (2023); Tekkesinoglu & Pudas (2024)
[b] Guo et al. (2018a); Fernando et al. (2019); Fusco et al. (2019); Marques-Silva et al. (2020); Nguyen & Martínez (2020)
[c] Jain & Wallace (2019); Wang et al. (2020b); Warnecke et al. (2020); Graziani et al. (2021); Rajbahadur et al. (2021); Krishna et al. (2022); Mercier et al. (2022)

Metric II.18: Significance Check

Desiderata: Fidelity
Explanation Type: FA, CE, (ExE, WBS, NLE) **References:** (15)
Samek et al. (2016); Adel et al. (2018); Chen et al. (2018a); Kim et al. (2018); Chen et al. (2019a); DeYoung et al. (2019); Gu et al. (2019); Wickramanayake et al. (2019); Nam et al. (2020); Hemamou et al. (2021); Park & Wallraven (2021); Pornprasit et al. (2021); Agarwal et al. (2022c); Hameed et al. (2022); Bommer et al. (2024)

Statistical significance testing is a common strategy for verifying whether an explanation is meaningfully different from random or naïve baseline explanations. This approach is frequently applied to FAs[a].
For CEs, statistical tests are typically used to assess whether the extracted concepts carry significantly more information than randomly sampled concepts. This is done by comparing average relevance or activation scores between the true and random concepts (Adel et al., 2018; Kim et al., 2018).
In a more advanced variant, Hemamou et al. (2021) train a classifier to distinguish between real and synthetic (random) explanations. High accuracy in this task indicates that the generated explanantia contain statistically meaningful structure.
While commonly reported for FA and CE, the general idea of statistical significance testing could, in principle, be adapted to other explanation types as well. However, doing so may require explanation-specific reformulations.

[a] Samek et al. (2016); Chen et al. (2018a; 2019a); DeYoung et al. (2019); Gu et al. (2019); Wickramanayake et al. (2019); Nam et al. (2020); Hemamou et al. (2021); Park & Wallraven (2021); Pornprasit et al. (2021); Agarwal et al. (2022c); Hameed et al. (2022); Bommer et al. (2024)

Metric II.19: (Counter-)Factual Relevance

Desiderata: Fidelity
Explanation Type: ExE **References:** (1)
Liu et al. (2021b)

This metric applies to explanations that generate both a factual and counterfactual explanans, aiming to evaluate how well they reflect and contrast the model's reasoning.
Originally proposed by Liu et al. (2021b) for the graph domain, requiring both explanantia being subgraphs, it can be generalized to other input types.
The method computes the model's output for the factual and counterfactual explanantia and compares them to the original prediction using the negative symmetric Kullback-Leibler divergence. The difference between these two scores is normalized by the distance between the factual and counterfactual explanantia. A high normalized score indicates that both explanans are informative: the factual preserves the original reasoning, while the counterfactual shifts the decision appropriately.

Metric II.20: Prediction Validity

Desiderata: Fidelity
Explanation Type: ExE **References:** (18)
Wachter et al. (2017); Karlsson et al. (2018); Dhurandhar et al. (2019); Guidotti et al. (2019); Mahajan et al. (2019); Dandl et al. (2020); Le et al. (2020); Molnar (2020); Mothilal et al. (2020); Nguyen & Martínez (2020); Pedapati et al. (2020); Pawelczyk et al. (2021); Rasouli & Yu (2021); Ma et al. (2022); Tan et al. (2022); Vermeire et al. (2022); Guidotti (2024); Verma et al. (2024)

By definition, a counterfactual must result in a different prediction from the original input. For untargeted counterfactuals, this simply requires $\hat{y}_z \neq \hat{y}_x$, while targeted counterfactuals must satisfy $\hat{y}_z = y_z^*$ (Wachter et al., 2017; Molnar, 2020; Guidotti, 2024).
Most authors assess the fraction of generated counterfactuals that meet this condition[a]. To capture more nuance, authors also propose measuring the model's confidence in the target class for the counterfactual (Rasouli & Yu, 2021), or using a continuous loss between the predicted and target class probabilities, such as the L_1 distance (Dandl et al., 2020).
For factual explanations, the class should remain unchanged, i.e., $\hat{y}_z = \hat{y}_x$, which can also be verified either binary (Dhurandhar et al., 2019) or using loss-based similarity measures (Nguyen & Martínez, 2020).

[a]Karlsson et al. (2018); Dhurandhar et al. (2019); Guidotti et al. (2019); Mahajan et al. (2019); Le et al. (2020); Mothilal et al. (2020); Pedapati et al. (2020); Pawelczyk et al. (2021); Ma et al. (2022); Tan et al. (2022); Vermeire et al. (2022); Verma et al. (2024)

Metric II.21: Sufficiency

Desiderata: Fidelity
Explanation Type: CE **References:** (2)
Yeh et al. (2020); Dasgupta et al. (2022)

An explanation should be complete enough to serve as a sufficient reason for a given model output. Ideally, the explanans alone should allow us to predict the outcome. This property reflects the completeness of the explanans and can be assessed in different ways.
Yeh et al. (2020) train a secondary model that maps the explanans back to the black-box's activation space. The predictive performance using the mapped explanans indicates how much information the explanation preserves about the original model's decision process.
Alternatively, Dasgupta et al. (2022) evaluate whether explanations are sufficient for consistent outcomes: Given an explanans e_0 for an input x_0, we identify other instances whose explanantia are equivalent (or sufficiently similar) to e_0 and compute the fraction that shares the same model prediction as x_0. A higher agreement indicates a more sufficient explanation.

Metric II.22: Output Faithfulness

Desiderata: Fidelity
Explanation Type: WBS **References:** (34)
Andrews et al. (1995); Craven & Shavlik (1995); Stefanowski & Vanderpooten (2001); Barakat et al. (2010); Augasta & Kathirvalavakumar (2012); Zilke et al. (2016); Bastani et al. (2017); Krishnan & Wu (2017); Lakkaraju et al. (2017); Guo et al. (2018b); Laugel et al. (2018); Peake & Wang (2018); Plumb et al. (2018); Tan et al. (2018); Wu et al. (2018a); Zhang et al. (2018b); Chen et al. (2019a); Guidotti et al. (2019); Kanehira & Harada (2019); Lakkaraju et al. (2019); Zhou et al. (2019); Anders et al. (2020); Hatwell et al. (2020); Lakkaraju et al. (2020); Panigutti et al. (2020); Pedapati et al. (2020); Rajapaksha et al. (2020); Rawal & Lakkaraju (2020); Amparore et al. (2021); Chen et al. (2021); Moradi & Samwald (2021); Pornprasit et al. (2021); Bayrak & Bach (2023b); Bo et al. (2024)

A surrogate model should closely mimic the behavior of the original black-box model. Therefore, a common evaluation strategy is to compare the outputs of the surrogate and black-box models using a similarity or performance measure.
This can be assessed with standard measures such as accuracy, F_1-score, MSE, or L_p distances[a], or with similarity measures such as SSIM, Pearson correlation, or KL divergence (Chen et al., 2019a; Anders et al., 2020). Arbitrary loss functions may also be used (Amparore et al., 2021).
For imbalanced datasets, Moradi & Samwald (2021) calculate fidelity per class, using either the true labels or the black-box predictions as reference. For individual rules, fidelity may be expressed as the rule's precision (Stefanowski & Vanderpooten, 2001; Hatwell et al., 2020).
When evaluating local surrogates, *Local Output Fidelity* is defined by computing fidelity within a neighborhood of the input instance (Laugel et al., 2018; Plumb et al., 2018; Guidotti et al., 2019; Rajapaksha et al., 2020), or within a synthetic neighborhood (Panigutti et al., 2020; Pornprasit et al., 2021). For unseen data, Lakkaraju et al. (2020) validate explanations by comparing outputs to those of the nearest training instances.

[a]Andrews et al. (1995); Craven & Shavlik (1995); Barakat et al. (2010); Augasta & Kathirvalavakumar (2012); Zilke et al. (2016); Bastani et al. (2017); Krishnan & Wu (2017); Lakkaraju et al. (2017); Guo et al. (2018b); Laugel et al. (2018); Plumb et al. (2018); Tan et al. (2018); Zhang et al. (2018b); Guidotti et al. (2019); Kanehira & Harada (2019); Lakkaraju et al. (2019); Zhou et al. (2019); Anders et al. (2020); Lakkaraju et al. (2020); Panigutti et al. (2020); Pornprasit et al. (2021); Bayrak & Bach (2023b)

Metric II.23: Internal Faithfulness

Desiderata: WBS
Explanation Type: Fidelity **References:** (3)
Messalas et al. (2019); Anders et al. (2020); Amparore et al. (2021)

Since the WBS models can achieve similar predictive performance as the original black-box without relying on the same underlying reasoning (Messalas et al., 2019; Anders et al., 2020), it is essential to evaluate their internal fidelity, meaning the similarity in how both models justify their predictions. This can be achieved by comparing post-hoc explanations of the original and surrogate model for the same inputs. Using feature attribution methods (e.g., SHAP from Lundberg (2017)), a typical approach is to measure the average overlap of the top-k features between both models' explanantia (Messalas et al., 2019). Other similarity metrics and explanation types may also be used.
Alternatively, Amparore et al. (2021) compare counterfactuals generated from each model, treating their similarity as a proxy for the alignment of decision boundaries. This provides a structural view of how well the surrogate captures the black-box model's rationale beyond mere output agreement.

Metric II.24: Setup Consistency

Desiderata: Consistency
Explanation Type: FA, ExE, WBS, (CE, NLE) **References:** (11)
Bastani et al. (2017); Honegger (2018); Guidotti et al. (2019); Rajapaksha et al. (2020); Warnecke et al. (2020); Amparore et al. (2021); Graziani et al. (2021); Margot & Luta (2021); Velmurugan et al. (2021b); Dai et al. (2022); Vermeire et al. (2022)

This metric evaluates how consistent the explanantia remain when generated multiple times for the same input and model. This, however, is only a necessary consideration for nondeterministic explanation methods.
Some authors assess this axiomatically, by counting the fraction of identical explanantia produced across repeated runs (Honegger, 2018; Vermeire et al., 2022). Others calculate distances or similarity scores between different runs and report the aggregated variation, using general or task-specific metrics such as the variance of feature weights or feature presence[a].
Although most work focuses on local instance-wise explanations, the same principle applies to global explanations, where the similarity between globally constructed explanantia is measured (Bastani et al., 2017; Margot & Luta, 2021).
While we did not identify examples of this metric applied to every explanations type in the literature, the core idea is general and can easily be extended to any XAI algorithm.

[a] Guidotti et al. (2019); Rajapaksha et al. (2020); Warnecke et al. (2020); Amparore et al. (2021); Graziani et al. (2021); Velmurugan et al. (2021b); Dai et al. (2022)

Metric II.25: Hyperparameter Sensitivity

Desiderata: Consistency
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (6)
Chen et al. (2019a); Verma & Ganguly (2019); Bansal et al. (2020); Mishra et al. (2020); Sanchez-Lengeling et al. (2020); Graziani et al. (2021)

This metric evaluates how robust an explanation method is to changes in its configuration. Since many XAI methods rely on hyperparameters, high sensitivity may indicate instability, making tuning more difficult and reducing user trust in the method (Bansal et al., 2020).
A common approach is to generate explanantia for the same input across different hyperparameter settings and measure the similarity between them[a]. Alternatively, Chen et al. (2019a) propose to compare the stability of performance metrics such as fidelity over varying hyperparameters. Beyond

identifying general sensitivity, Graziani et al. (2021) use this technique to guide hyperparameter selection: by progressively adjusting hyperparameters and tracking when the generated explanantia converge, one can identify stable regions in the hyperparameter space.
Although most work focuses on FAs, the general idea is applicable to any method with tunable hyperparameters.

[a] Verma & Ganguly (2019); Bansal et al. (2020); Mishra et al. (2020); Sanchez-Lengeling et al. (2020); Graziani et al. (2021)

Metric II.26: Execution Time

Desiderata: Efficiency
Explanation Type: FA, ExE, WBS, (CE, NLE) **References:** (34)
Ross et al. (2017); Ribeiro et al. (2018); Zhang et al. (2018c); Chen et al. (2019a); Cheng et al. (2019); Fusco et al. (2019); Ignatiev et al. (2019); Shakerin & Gupta (2019); Slack et al. (2019); Topin & Veloso (2019); Albini et al. (2020); Guo et al. (2020); Marques-Silva et al. (2020); Rajapaksha et al. (2020); Ramon et al. (2020); Warnecke et al. (2020); Abrate & Bonchi (2021); Bajaj et al. (2021); Faber et al. (2021); Lin et al. (2021); Malik et al. (2021); Pawelczyk et al. (2021); Rasouli & Yu (2021); Van Looveren & Klaise (2021); Wang et al. (2021); Amoukou et al. (2022); Belaid et al. (2022); Ma et al. (2022); Mercier et al. (2022); Vermeire et al. (2022); Bayrak & Bach (2023a); Brandt et al. (2023); Jin et al. (2023); Verma et al. (2024)

Most authors measure the time required to generate an explanans for a fixed input, model, and dataset, on specified hardware[a].
In addition to empirical runtime, some authors analyze algorithmic complexity using $\mathcal{O}$-notation to characterize the worst-case or average-case number of steps required to run an explanation[b]. Chen et al. (2019a) further evaluate parallelizability to understand potential runtime improvements through hardware acceleration or distributed computation.

[a] Ross et al. (2017); Ribeiro et al. (2018); Zhang et al. (2018c); Cheng et al. (2019); Fusco et al. (2019); Ignatiev et al. (2019); Shakerin & Gupta (2019); Guo et al. (2020); Marques-Silva et al. (2020); Rajapaksha et al. (2020); Ramon et al. (2020); Warnecke et al. (2020); Bajaj et al. (2021); Faber et al. (2021); Lin et al. (2021); Malik et al. (2021); Pawelczyk et al. (2021); Rasouli & Yu (2021); Van Looveren & Klaise (2021); Wang et al. (2021); Amoukou et al. (2022); Belaid et al. (2022); Ma et al. (2022); Mercier et al. (2022); Vermeire et al. (2022); Bayrak & Bach (2023a); Brandt et al. (2023); Jin et al. (2023); Verma et al. (2024)
[b] Slack et al. (2019); Topin & Veloso (2019); Albini et al. (2020); Abrate & Bonchi (2021); Malik et al. (2021); Van Looveren & Klaise (2021)

B.3.III Input Intervention

The metrics listed in this group all require changes to the input, which is then presented to the model again and potentially explained once more.

Metric III.27: Unguided Perturbation Fidelity

Desiderata: Fidelity
Explanation Type: FA **References:** (10)
Ancona et al. (2017); Shrikumar et al. (2017); Sundararajan et al. (2017); Alvarez-Melis & Jaakkola (2018b); Arya et al. (2019); Cheng et al. (2019); Yeh et al. (2019); Zhang et al. (2019b); Bhatt et al. (2020); Elkhawaga et al. (2023)

Some of the most influential XAI metrics evaluate how well an explanans aligns with observed model behavior when the input is perturbed. Together, these metrics assess whether the explanation faithfully captures how the model responds to its inputs.
This includes *Sensitivity-n* from Ancona et al. (2017), *Infidelity* from Yeh et al. (2019), and *Faithfulness* presented by Alvarez-Melis & Jaakkola (2018b) and Arya et al. (2019). In these approaches, input perturbations are introduced either randomly across all features (Yeh et al., 2019), by zeroing

features individually or in small groups (Alvarez-Melis & Jaakkola, 2018b; Arya et al., 2019; Cheng et al., 2019; Bhatt et al., 2020), or by manipulating subsets of fixed size (Ancona et al., 2017). The model's change in prediction is then compared to the explanans, using different strategies: directly against the raw attribution vector, against a version scaled by perturbation magnitude, or by summing attributions of the changed features. The deviation is measured using standard metrics like Pearson correlation[a] or mean squared error (Yeh et al., 2019).
Complementing these are completeness-based approaches such as *Completeness* (Sundararajan et al., 2017) and *Summation-to-Delta* (Shrikumar et al., 2017), which assume that the explanans must fully account for the model's behavior. These compare the difference in output between an instance and a baseline (e.g., a fully perturbed input) with the sum of the attributions across changed features. Ideally, the two should match exactly. For practical purposes, however, the relative deviation between attribution sum and output delta can be used as a softer criterion.

[a] Ancona et al. (2017); Alvarez-Melis & Jaakkola (2018b); Arya et al. (2019); Cheng et al. (2019); Bhatt et al. (2020)

Metric III.28: Guided Perturbation Fidelity

Desiderata: Fidelity
Explanation Type: FA,CE **References:** (75)
Bach et al. (2015); Samek et al. (2016); Ancona et al. (2017); Dabkowski & Gal (2017); Shrikumar et al. (2017); Chattopadhay et al. (2018); Chu et al. (2018); Guo et al. (2018a;b); Liu et al. (2018b); Nguyen (2018); Petsiuk (2018); Wang et al. (2018b); Yang et al. (2018a); Annasamy & Sycara (2019); Arya et al. (2019); Brahimi et al. (2019); DeYoung et al. (2019); Fong et al. (2019); Kanehira et al. (2019); Kapishnikov et al. (2019); Lin et al. (2019); Pope et al. (2019); Schlegel et al. (2019); Serrano & Smith (2019); Wagner et al. (2019); Yuan et al. (2019); Ayush et al. (2020); Brunke et al. (2020); Carton et al. (2020); Chen et al. (2020); Cong et al. (2020); Fan et al. (2020); Hsieh et al. (2020); Nguyen et al. (2020); Pan et al. (2020b); Rieger & Hansen (2020); Schlegel et al. (2020); Schulz et al. (2020); Singh et al. (2020); Wang et al. (2020a); Warnecke et al. (2020); El Shawi et al. (2021); Ge et al. (2021); Jethani et al. (2021); Jung & Oh (2021); Liu et al. (2021b); Luss et al. (2021); Poppi et al. (2021); Singh et al. (2021); Situ et al. (2021); Sun et al. (2021); Velmurugan et al. (2021a;b); Vu et al. (2021); Wang et al. (2021); Yin et al. (2021); Albini et al. (2022); Atanasova et al. (2022); Dai et al. (2022); De Silva et al. (2022); Funke et al. (2022); Hameed et al. (2022); Müller et al. (2022); Ngai & Rudzicz (2022); Rong et al. (2022); Tan et al. (2022); Veerappa et al. (2022); Šimić et al. (2022); Zou et al. (2022); Agarwal et al. (2023); Alangari et al. (2023b); Jin et al. (2023); Schlegel & Keim (2023); Awal & Roy (2024)

A common and versatile strategy for evaluating explanation quality is to assess how perturbing features based on the explanans affects model predictions. This approach captures both the necessity (correctness) and sufficiency (completeness) of features highlighted by an explanans (Carton et al., 2020; DeYoung et al., 2019; Alangari et al., 2023b). Details on perturbation strategies are outlined in Subsection B.2.1.
Perturbation Scope: Perturbations can be applied either in a fixed-size or iterative manner. Fixed-size perturbations change a predefined amount of information at once[a]. The amount of removed features can be selected as a fixed number[b] or based on a threshold over cumulative importance scores[c]. Iterative perturbation gradually removes features one by one or in batches[d].
Perturbation Order: The order in which features are perturbed is critical. Most relevant first (MoRF) evaluates correctness: removing the most important features should lead to a sharp performance drop if the explanation correctly highlights necessary features[e]. Least relevant first (LeRF) tests completeness: performance should remain high if unimportant features are removed first[f]. Alternatively, some authors reverse the entire process by starting from a blank input and adding features incrementally[g], which is functionally equivalent, with MoRF addition corresponding to LeRF deletion, and vice-versa (for an illustration see Figure 7).
Measure Score: The performance change is measured using various scoring mechanisms. These include raw prediction difference $\theta(x, \hat{y}) - \theta(\dot{x}, \hat{y})$[h], normalized or relative variants[i], or evaluation of loss and accuracy metrics[j]. Some works use statistical distances such as Kullback-Leibler divergence

(Liu et al., 2021b; Agarwal et al., 2023), Kendall's Tau (Singh et al., 2020; 2021), or Pearson correlation (Poppi et al., 2021).

Aggregation: For iterative perturbations, the results can be aggregated as the average performance change (Samek et al., 2016; Ancona et al., 2017), the area over the perturbation curve (AOPC) (Samek et al., 2016; Brahimi et al., 2019; DeYoung et al., 2019; Rieger & Hansen, 2020), or the area under it (AUPC) (Petsiuk, 2018; Ngai & Rudzicz, 2022; Šimić et al., 2022; Jin et al., 2023). The area between MoRF and LeRF (ABPC) may also be computed (Nguyen et al., 2020; Schulz et al., 2020), optionally with decay weighting to emphasize early changes (Šimić et al., 2022).

Baselines: To contextualize results, explanation can be compared against various baselines, including random feature selections[k], zero-inputs (Schulz et al., 2020), or naïve explanation such as edge detectors (Hooker et al., 2019). Advanced setups may compare against models trained on random labels or with inserted irrelevant features to bound the quality of explanation (Hameed et al., 2022).

Variants:

- Evaluating explanantia across all classes per instance, not only the predicted class (Pope et al., 2019; Awal & Roy, 2024).
- Sanity checks verifying whether the perturbation curve behaves monotonically (Arya et al., 2019; Fong et al., 2019; Luss et al., 2021), or whether MoRF always performs worse than LeRF (Šimić et al., 2022).
- Normalizing the performance drop by the input difference to mitigate distribution shift: $\frac{\delta\left(\theta(x),\theta(\dot{x})\right)}{\delta(x,\dot{x})}$ (Ge et al., 2021; Schlegel & Keim, 2023).
- Plotting performance over remaining entropy rather than number of perturbed features (Kapishnikov et al., 2019).
- Training with randomized feature masking to improve model stability, although this may reduce causality fidelity (Jethani et al., 2021).

Alternatives: Further, we may change the perspective through alternative formulations:

- Counting the minimum number of features needed to change the model's prediction (in MoRF deletion) (Nguyen, 2018) or using differences between class-specific explanantia to guide perturbations (Shrikumar et al., 2017).
- Treating perturbed inputs as counterfactuals and applying metrics such as Minimality (see Metric 4) (Fan et al., 2020; Ge et al., 2021; Albini et al., 2022).
- Replacing fixed perturbations with adversarial optimization over k features to evaluate minimum change necessary for altering predictions (Hsieh et al., 2020; Vu et al., 2021).

The approach may be extended to CEs by either mapping concepts to features (e.g., concept-based saliency maps by Lucieri et al. (2020)), or perturbing internal activations at the concept layer (El Shawi et al., 2021).

[a] Guo et al. (2018b); DeYoung et al. (2019); Warnecke et al. (2020); Jethani et al. (2021); Jung & Oh (2021); Velmurugan et al. (2021a;b); Wang et al. (2021); De Silva et al. (2022); Agarwal et al. (2023)

[b] Bach et al. (2015); Samek et al. (2016); Ancona et al. (2017); Chu et al. (2018); Schlegel et al. (2019); Nguyen et al. (2020); Alangari et al. (2023b); Schlegel & Keim (2023)

[c] Brahimi et al. (2019); Pope et al. (2019); Schlegel et al. (2019); Ngai & Rudzicz (2022); Schlegel & Keim (2023)

[d] Bach et al. (2015); Samek et al. (2016); Ancona et al. (2017); Petsiuk (2018); DeYoung et al. (2019); Rieger & Hansen (2020); Jin et al. (2023)

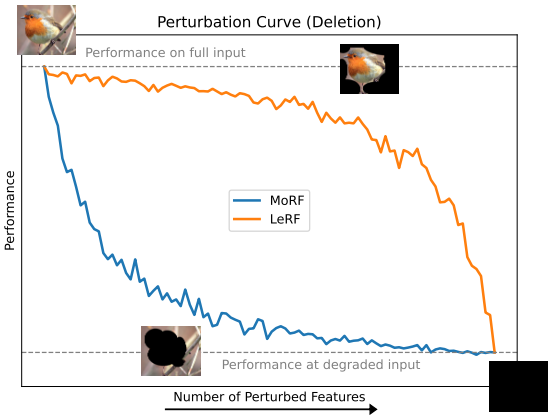
[e] Bach et al. (2015); Samek et al. (2016); Chu et al. (2018); DeYoung et al. (2019); Pope et al. (2019); Schlegel et al. (2019); Carton et al. (2020); Rieger & Hansen (2020); Ngai & Rudzicz (2022); Alangari et al. (2023b); Jin et al. (2023); Schlegel & Keim (2023)

[f] Bach et al. (2015); Ancona et al. (2017); Dabkowski & Gal (2017); Guo et al. (2018b); Annasamy & Sycara (2019); DeYoung et al. (2019); Wagner et al. (2019); Carton et al. (2020); Singh et al. (2020); Wang et al. (2020a); Jethani et al. (2021); Liu et al. (2021b); Singh et al. (2021); Wang et al. (2021); Dai et al. (2022); De Silva et al. (2022); Alangari et al. (2023b)

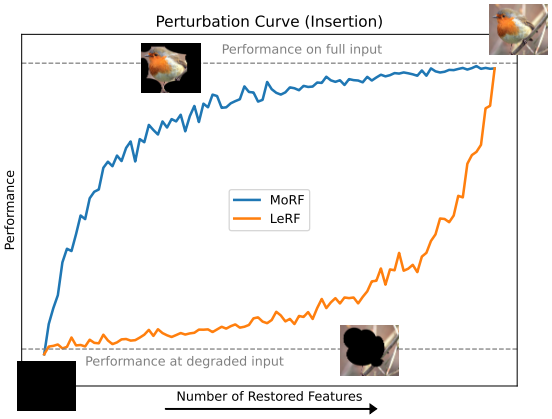
[g] Petsiuk (2018); Arya et al. (2019); Kapishnikov et al. (2019); Luss et al. (2021); Situ et al. (2021); Sun et al. (2021); Atanasova et al. (2022); Funke et al. (2022); Tan et al. (2022)

[h] Bach et al. (2015); Chu et al. (2018); DeYoung et al. (2019); Kapishnikov et al. (2019); Chen et al. (2020); Cong et al. (2020); Schulz et al. (2020); Ngai & Rudzicz (2022)

[i]Chattopadhay et al. (2018); Kapishnikov et al. (2019); Schulz et al. (2020); Jung & Oh (2021); Velmurugan et al. (2021a;b); Ngai & Rudzicz (2022)
[j]Bach et al. (2015); Chu et al. (2018); Kapishnikov et al. (2019); Lin et al. (2019); Schlegel et al. (2019); Serrano & Smith (2019); Wagner et al. (2019); Warnecke et al. (2020); Sun et al. (2021); Wang et al. (2021); De Silva et al. (2022); Schlegel & Keim (2023)
[k]Samek et al. (2016); Schlegel et al. (2019); Schlegel & Keim (2023); Serrano & Smith (2019); Jin et al. (2023)



(a) Deletion perturbs features iteratively.



(b) Insertion restores features iteratively.

Figure 7: Examples for ideal Perturbation Curves to evaluate feature attributions. Insertion and Deletion approaches are equivalent when switching MoRF for LeRF ordering and vice versa. Deletion MoRF (Insertion LeRF) can be used to evaluate correctness, and Insertion MoRF (Deletion LeRF) to evaluate completeness.

Metric III.29: Counterfactuability

Desiderata: Fidelity
Explanation Type: WBS
References: (1)
Pornprasit et al. (2021)

To assess the expressiveness and impact of rule-based explanations (either generated directly or extracted from surrogate models such as trees), we can use them to guide the generation of counterfactual perturbations. This evaluates whether the rules have predictive leverage and reflect real decision logic, rather than being purely descriptive or spurious.

Specifically, given an input instance that is covered by a rule, Pornprasit et al. (2021) perturb the input such that the rule conditions are violated. If the rule truly captures important rationale, breaking it should influence the black-box model's prediction. This can be quantified in two ways:

- By counting the number of perturbed instances that change the predicted class label.
- By measuring the aggregate change in predicted class probability before and after perturbation.

Metric III.30: Prediction Neighborhood Continuity

Desiderata: Fidelity, Continuity
Explanation Type: WBS **References:** (1)
Lakkaraju et al. (2020)

A WBS is only useful if its behavior remains consistent with the black-box even under slight perturbations to the input.
To assess this, Lakkaraju et al. (2020) propose measuring the difference in Output Faithfulness (see Metric 22) between the original and perturbed inputs. Specifically, for each input a perturbed variant is created (e.g., through noise or adversarial modification). The metric then compares the agreement between the black-box model and the surrogate model before and after the perturbations
A lower difference indicates a more robust surrogate, as it preserves faithfulness across perturbations. High differences may signal that the surrogate captures only superficial model behavior or overfits to specific input patterns.

Metric III.31: Quantification of Unexplainable Features

Desiderata: Continuity, (Fidelity)
Explanation Type: FA, (CE) **References:** (2)
Zhang et al. (2019a); Chen et al. (2022)

To assess the continuity of explanations, Zhang et al. (2019a) perturb the unimportant features of an input based on the explanans, and then generate a new one for the perturbed input. The similarity between the original and perturbed explanantia serves as a measure of continuity: a high similarity suggests that irrelevant changes do not affect the explanation, indicating robustness. In addition, it implies a high completeness, as relevant features must have been captured well enough to maintain the model's behavior despite noise.
A special case of this approach is the Attack Capture Rate proposed by Chen et al. (2022), which applies to FAs in NLP (particularly rationale extraction). Here, insertion attacks introduce distractor phrases to the input, which ideally should not influence the rationale. The metric measures how often the inserted tokens appear in the extracted explanans. This variant relies on artificial attacks, which limits its generalizability beyond NLP.
This metric may be extended to CE by perturbing the concept layer corresponding to unimportant concepts, or by mapping concepts back to input features via concept-based localization (e.g., see Lucieri et al. (2020)).

Metric III.32: Neighborhood Continuity

Desiderata: Continuity
Explanation Type: FA, ExE, WBS, NLE, (CE) **References:** (19)
Alvarez-Melis & Jaakkola (2018a;b); Chu et al. (2018); Honegger (2018); Yeh et al. (2019); Zhang et al. (2019a); Artelt & Hammer (2020); Fan et al. (2020); Lakkaraju et al. (2020); Artelt et al. (2021); Bajaj et al. (2021); Situ et al. (2021);

Yin et al. (2021); Agarwal et al. (2022a); Atanasova et al. (2022); Fouladgar et al. (2022); Agarwal et al. (2023); Bayrak & Bach (2023a); Tekkesinoglu & Pudas (2024)

Neighborhood Continuity evaluates whether explanantia remain similar for similar explananda. The underlying assumption is that small changes in the input should not lead to disproportionately large differences in the output, thereby increasing user trust through Continuity.
The main distinction between proposed metrics lies in how "similar instances" are defined. Some classic metrics include *Local Lipschitz Stability* from Alvarez-Melis & Jaakkola (2018a;b), *Sensitivity* from Yeh et al. (2019), and *RIS/RRS/ROS* from Agarwal et al. (2022a).
The **choice of neighborhood** depends on the domain and explanation use case. Some approaches define neighborhoods using fixed-radius input distances or k-nearest neighbors (Chu et al., 2018; Situ et al., 2021; Fouladgar et al., 2022; Tekkesinoglu & Pudas, 2024), while others restrict similarity to instances sharing the same predicted label (Honegger, 2018; Fan et al., 2020).
To **generate similar inputs**, perturbation-based strategies are often used (see Subsection B.2.1). Perturbations may be bounded in magnitude[a] or derived from domain-specific semantics (Zhang et al., 2019a; Yin et al., 2021; Fouladgar et al., 2022). Some metrics require that perturbed inputs preserve the original prediction $\hat{y}_x$ (Alvarez-Melis & Jaakkola, 2018a;b; Agarwal et al., 2022a), or maintain similar logits $\theta(x)$ (Agarwal et al., 2023).
Explanans similarity is typically calculated using distance or correlation-based metrics (see Subsection B.2.3) and may be normalized by input distance (Alvarez-Melis & Jaakkola, 2018a;b; Agarwal et al., 2022a), or based on distances in model activation space (Agarwal et al., 2022a; 2023).
Variations: Fan et al. (2020) compare similarity for nearby vs. distant inputs , while Zhang et al. (2019a) restrict comparisons to unchanged features to reduce noise from perturbations.
Although originally proposed for FAs, this concept is broadly transferable to other explanation types. For instance, Lakkaraju et al. (2020) evaluate WBSs by comparing surrogate models trained on perturbed data, using model-specific similarity measures such as coefficient mismatch for linear models.

[a] Alvarez-Melis & Jaakkola (2018a;b); Yeh et al. (2019); Bajaj et al. (2021); Agarwal et al. (2022a); Atanasova et al. (2022)

Metric III.33: Adversarial Input Resilience

Desiderata: Continuity
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (10)
Singh et al. (2018); Wang et al. (2018b); Chen et al. (2019b); Dombrowski et al. (2019); Ghorbani et al. (2019); Subramanya et al. (2019); Boopathy et al. (2020); Kuppa & Le-Khac (2020); Zhang et al. (2020); Huang et al. (2023a)

Explanations can be vulnerable to Adversarial Attacks (AA), where the goal is to manipulate either the explanans (Ghorbani et al., 2019) or prediction (Singh et al., 2018; Wang et al., 2018b; Subramanya et al., 2019). More sophisticated approaches aim to manipulate one, while additionally restricting changes to the other. Two main types exist:

- **Explanans manipulation**: The explanans is altered while the prediction remains fixed (Dombrowski et al., 2019; Boopathy et al., 2020; Kuppa & Le-Khac, 2020; Huang et al., 2023a).
- **Prediction manipulation**: The prediction changes while the explanans remains similar (Zhang et al., 2020; Huang et al., 2023a).

Evaluation typically measures the distance between the adversarial output (explanans or prediction) and either its original or targeted counterpart. The input change is usually bounded. Performance is reported either as aggregated distance metrics (Dombrowski et al., 2019; Ghorbani et al., 2019; Zhang et al., 2020; Huang et al., 2023a) or as attack success rates based on predefined thresholds (Kuppa & Le-Khac, 2020; Zhang et al., 2020; Huang et al., 2023a).

While the reported authors focus on FAs, the approach can be easily adapted to other explanation types by selecting appropriate similarity measures.

B.3.IV Model Intervention

For some categories of metrics, it is necessary to change the underlying model, e.g., through retraining or weight randomization.

Metric IV.34: Model Parameter Randomization Test

Desiderata: Fidelity
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (5)
Adebayo et al. (2018); Kindermans et al. (2019); Binder et al. (2023); Bommer et al. (2024); Hedström et al. (2024)

To verify that explanations are truly reflective of the black-box model's learned reasoning, the model's internal parameters are systematically randomized, and the resulting changes in explanantia are analyzed. The rationale is that if an explanation remains unchanged under randomization, it is likely generic and not informative of the model's decision logic (Adebayo et al., 2018).
Most commonly, the weights of a neural network are randomized either entirely, by layer, or iteratively in top-down or bottom-up order. The similarity between the original and randomized explanantia is then computed (Adebayo et al., 2018; Hedström et al., 2024). Similarity may be evaluated using SSIM and correlation for heatmaps (Adebayo et al., 2018; Binder et al., 2023; Bommer et al., 2024), or any suitable similarity metric (see Subsection B.2.3). Hedström et al. (2024) additionally average the results across noisy inputs (e.g., via input perturbations) to stabilize local explanations.
For gradient-based methods, Sixt et al. (2020) propose inserting a random activation vector at a specific layer, to break the causal connection without modifying the network weights.
A complementary strategy proposed by Kindermans et al. (2019) keeps the weights fixed but applies a controlled shift to the input distribution, adapting the input-layer biases such that internal computations and outputs remain unchanged. Since the model behavior is invariant by design, any change in the explanans indicates unwanted sensitivity. This shift-invariance test can be quantified using standard similarity metrics between pre- and post-shift explanantia.
Although originally proposed for FAs, the approach is applicable to any explanation type for which suitable similarity metrics can be defined.

Metric IV.35: Data Randomization Test

Desiderata: Fidelity
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (2)
Adebayo et al. (2018); Sanchez-Lengeling et al. (2020)

Explanations should highlight meaningful structures in the data, not artifacts of memorization. To verify this, the training data labels are randomized, forcing the model to fit noise rather than learn semantically relevant features (Adebayo et al., 2018). Label randomization can be applied to the full training set (Adebayo et al., 2018) or a subset only (Sanchez-Lengeling et al., 2020).
This test compares explanantia generated by a model trained on the randomized data to those from a model trained on correctly labeled data. A strong explanation method should yield low similarity between the two, as the latter explanantia reflect meaningful decision features, while the former do not. Similarity can be computed using SSIM, correlation (Adebayo et al., 2018), rank-based metrics such as Kendall's Tau (Sanchez-Lengeling et al., 2020), or other suitable measures (see Subsection B.2.3).
Although only reported for FAs, this approach can be extended to any explanation type, provided that a suitable similarity measure between explanantia is available.

Metric IV.36: Retrained Model Evaluation

Desiderata: Fidelity
Explanation Type: FA **References:** (10)
Guo et al. (2019); Hooker et al. (2019); Cheng et al. (2020); Han et al. (2020); Schiller et al. (2020); Hemamou et al. (2021); Shah et al. (2021); Li et al. (2022); Khalane et al. (2023); Raval et al. (2023)

A key limitation of perturbation-based evaluation methods lies in their potential to introduce distribution shifts though the alteration of features. To mitigate this, two related strategies retrain models on the perturbed datasets.
The first approach is most famously known as Remove and Retrain (ROAR) by Hooker et al. (2019), with related variants proposed by Han et al. (2020) and Shah et al. (2021). It builds on the perturbation strategies of Metric 28, but ROAR removes features (e.g., the most important ones according to the explanans) from the training data and then retrains the model from scratch on the altered dataset. Performance degradation of this retrained model, compared to the original model, is then used to infer the quality of the explanation: a faithful explanans should identify features whose removal severely affects performance.
The second strategy leverages explanations for knowledge distillation. Here, various authors train a new model on a dataset reduced to only the features deemed important by the explanation[a].
Evaluation follows the general structure of Metric 28, measuring how much predictive performance is retained when only the relevant features are used.
While both ROAR and distillation-based setups are valuable for benchmarking explanation methods in a more robust setting, they do not directly assess the original explanans for a specific instance. Instead, they evaluate the utility of the explanation method by measuring the average informativeness of the features it selects.

[a] Guo et al. (2019); Cheng et al. (2020); Schiller et al. (2020); Hemamou et al. (2021); Li et al. (2022); Khalane et al. (2023); Raval et al. (2023)

Metric IV.37: Influence Fidelity

Desiderata: Fidelity
Explanation Type: ExE **References:** (2)
Krishnan & Wu (2017); Guo et al. (2020)

To evaluate the fidelity of ExEs, the model is retrained on a dataset modified according to the explanation. The effect of these modifications on model predictions is used to assess the explanatory quality.
Guo et al. (2020) remove the identified influential training points from the dataset, retrain the model, and measure the change in loss for the explanandum. If the removed instances were truly helpful, model performance should degrade. Conversely, if they were misleading, removal should lead to an improvement. Thus, a greater change in loss signals a more correct and complete set of influential instances.
In an alternative approach, Krishnan & Wu (2017) retain the identified influential instances but randomly flips the labels of all remaining training examples before retraining. If the explanation captures all relevant information, the prediction should remain stable. Hence, lower prediction variance after retraining indicates a more complete and accurate explanans.

Metric IV.38: Normalized Movement Rate

Desiderata: Continuity
Explanation Type: FA **References:** (2)
Salih et al. (2022; 2024b)

To assess the robustness of FAs in the presence of collinear or redundant features, Salih et al. (2022) introduce a metric that evaluates the stability of feature rankings as the most important features are iteratively removed and the model is retrained. This procedure mirrors the setup of the Remove and Retrain paradigm (Metric 36), but shifts the focus from prediction performance to the behavior of the explanation itself.
After each retraining step, the explanans is recomputed, and the ranks of the remaining features are compared to the previous ones. A large shift in ranking indicates that redundant or weakly relevant features have taken over the role of more informative ones, suggesting that the attribution method lacks robustness in the face of redundancy and may not reflect the true rationale behind the model's prediction.
To address this issue, Salih et al. (2024b) propose the Modified Informative Position (MIP), which aims to stabilize the interpretation by providing a more resilient ranking structure across iterative retraining steps.

Metric IV.39: Adversarial Model Resilience

Desiderata: Continuity
Explanation Type: FA, (ExE, CE, WBS, NLE) **References:** (4)
Heo et al. (2019); Pruthi et al. (2019); Viering et al. (2019); Dimanov et al. (2020)

Explanation methods should depend meaningfully on the internal parameters of the black-box model. However, this sensitivity can be exploited to adversarially manipulate them. Specifically, small but carefully chosen changes to the model's weights can alter the generated explanations without significantly changing predictions. This manipulation can serve to obfuscate undesirable behaviors or bias within a model.
To assess the robustness of explanation methods against such attacks, various strategies perturb the model parameters and observe the resulting changes in explanantia. These attacks can be:

- *Targeted*, where the modified model is encouraged to produce explanantia similar to a predefined target (Heo et al., 2019; Viering et al., 2019).
- *Untargeted*, where the goal is to produce explanantia that differ substantially from the original ones (Heo et al., 2019; Pruthi et al., 2019; Dimanov et al., 2020).

To preserve prediction behavior, constraints may be imposed on the model modification. Dimanov et al. (2020) bound the change in prediction scores , but other restrictions are thinkable, such as limiting the weight difference norm to ensure the model remains close to the original.
The success of the manipulation is measured depending on the attack setup. For targeted attacks, its the similarity between the manipulated explanans and the predefined target (Viering et al., 2019). Untargeted attacks are evaluated using the dissimilarity between the manipulated and original explanans (Dimanov et al., 2020; Pruthi et al., 2019).
All approaches were originally introduced for FA, but by applying suitable similarity measures (see Subsection B.2.3), they can be extended to other explanation types as well.

B.3.V A Priori Constrained

The final group of metrics is those that have additional requirements to the desideratum, namely the necessity for specific annotations in the dataset or using a specific type of models.

Metric V.40: GT Dataset Evaluation

Desiderata: Plausibility, Fidelity
Explanation Type: FA, CE, NLE **References:** (119)
Human-Annotated Dataset (69): Simonyan et al. (2013); Cao et al. (2015); Lapuschkin et al. (2016); Zhou et al. (2016); Das et al. (2017); Fong & Vedaldi (2017); Selvaraju et al. (2017); Bargal et al. (2018); Baumgartner et al. (2018); Camburu et al. (2018); Chen et al. (2018a); Chuang et al. (2018); Jha et al. (2018); Liu et al. (2018a); Park et al. (2018); Poerner et al. (2018); Wang et al. (2018a); Wu & Mooney (2018); Zhang et al. (2018a); Bastings et al. (2019); Chen et al. (2019d;c); DeYoung et al. (2019); Fong et al. (2019); Kanehira & Harada (2019); Kanehira et al. (2019); Mitsuhara et al. (2019); Puri et al. (2019); Rajani et al. (2019); Shu et al. (2019); Sydorova et al. (2019); Taghanaki et al. (2019); Trokielewicz et al. (2019); Verma & Ganguly (2019); Wang & Nvasconcelos (2019); Wickramanayake et al. (2019); Zeng et al. (2019); Bass et al. (2020); Cheng et al. (2020); Li et al. (2020a); Liu et al. (2020); Nam et al. (2020); Pan et al. (2020a); Rio-Torto et al. (2020); Schulz et al. (2020); Subramanian et al. (2020); Sun et al. (2020); Wang & Vasconcelos (2020); Xu et al. (2020a;b); Bany Muhammad & Yeasin (2021); Barnett et al. (2021); Bykov et al. (2021); Jang & Lukasiewicz (2021); Joshi et al. (2021); Mathew et al. (2021); Nguyen et al. (2021); Wiegreffe & Marasović (2021); Asokan et al. (2022); Cui et al. (2022); Du et al. (2022); Mücke & Pfahler (2022); Theiner et al. (2022); Nematzadeh et al. (2023); Rasmussen et al. (2023); Ribeiro et al. (2023); Tritscher et al. (2023); Atanasova (2024b); Saifullah et al. (2024)
Synthetic Dataset (50): Cortez & Embrechts (2013); Chen et al. (2017); Oramas et al. (2017); Ross et al. (2017); Chen et al. (2018b); Kim et al. (2018); Mascharka et al. (2018); Yang et al. (2018b); Antwarg et al. (2019); Arras et al. (2019); Camburu et al. (2019); Ismail et al. (2019); Jia et al. (2019); Lin et al. (2019); Subramanya et al. (2019); Takeishi (2019); Yang & Kim (2019); Ying et al. (2019); Amiri et al. (2020); Ismail et al. (2020); Jia et al. (2020); Kohlbrenner et al. (2020); Lucieri et al. (2020); Luo et al. (2020); Nguyen & Martínez (2020); Tritscher et al. (2020); Wang et al. (2020a); Bohle et al. (2021); Faber et al. (2021); Kim et al. (2021); Lin et al. (2021); Liu et al. (2021a); Shah et al. (2021); Yalcin et al. (2021); Agarwal et al. (2022b); Amoukou et al. (2022); Arias-Duart et al. (2022); Arras et al. (2022); Fan et al. (2022); Khakzar et al. (2022); Rao et al. (2022); Tjoa & Guan (2022); Wilming et al. (2022); Zhou et al. (2022); Agarwal et al. (2023); Hesse et al. (2023); Miró-Nicolau et al. (2023); Sun et al. (2023); Ya et al. (2023); Rao et al. (2024)

To evaluate the quality of explanations, many authors propose comparing them to dataset-based ground truths. This strategy can follow two distinct paradigms, depending on how the ground truth is derived. If the rationale is uniquely defined through the data generation process, we may evaluate the explanation's **Fidelity**. If, however, the ground truth stems from human judgment or heuristic annotation, we evaluate the **Plausibility** of the explanation.
The former is feasible with synthetic datasets that encode precisely one explanatory rationale. The latter is more common in practice but inherently less reliable, as black-box models may learn spurious correlations as shown by Ribeiro et al. (2016) or identify valid rationales beyond those provided by humans (Mücke & Pfahler, 2022; Ya et al., 2023).
Synthetic Datasets are constructed such that the relationship between input features and labels is controlled and known. Ideally, these datasets admit only a single valid rationale. Although this assumption is not always met in practice[a]. The ground-truth explanantia may either be explicitly specified, e.g., by indicating the features responsible for prediction[b], or derived from the data-generation process itself, for instance using gradients or Shapley values[c]. Figure 8 presents selected examples.
Typical strategies for synthetic data include:

- Creating inputs from structured primitives (e.g., objects or shapes), where the target labels are deterministic functions of those primitives[d].
- Inserting backdoor triggers into inputs, which act as the decisive explanans for altered predictions[e].
- Performing adversarial attacks restricted to known feature subsets, which then serve as the explanatory evidence for prediction changes (Subramanya et al., 2019).
- Generating labels as noise-free functions over randomly sampled input features, commonly in tabular domains[f].
- Constructing mosaic images or input grids where the rationale corresponds to a localized sub-region of the input[g].

Advantages of synthetic datasets include reduced distribution shift when applying perturbation-based metrics (Metric 28) (Ismail et al., 2019; 2020; Hesse et al., 2023), support for concept-level manip-

ulations (Yang & Kim, 2019; Lin et al., 2021; Hesse et al., 2023), and the inclusion of meaningful counterfactuals or rationale variants (Yang & Kim, 2019).

Human-Annotated Datasets instead provide plausible rationales grounded in human intuition or derived from proxy labels. In some cases, annotators are explicitly asked to justify their decisions or to highlight which features they consider relevant for a particular prediction (Chen et al., 2019c; Xu et al., 2020b; Cui et al., 2022; Tritscher et al., 2023). In other settings, rationales are implicit, where existing annotations such as segmentation maps, bounding boxes, or other metadata are repurposed to approximate human reasoning[h]. These datasets are typically used to evaluate the plausibility of explanations rather than their fidelity, as the ground-truth provided reflects human expectations rather than the actual decision-making logic of the model.

Evaluation:

Across both paradigms, the central assumption is that accurate model predictions imply alignment with the intended rationale, thereby justifying a comparison between the generated explanantia and the ground-truth annotations. This comparison can be performed using any suitable similarity metric (see Subsection B.2.3). While FA evaluations often rely on feature-level scoring, CE and NLEs require domain-appropriate measures.

In the context of **FAs**, several dedicated evaluation strategies were reported. Precision and recall can be computed over truly important features to assess how well the explanation captures relevant inputs (Bastings et al., 2019). Other techniques evaluate the ranking of important and unimportant features (Chen et al., 2017; 2018b; Antwarg et al., 2019; Camburu et al., 2019), or quantify the symmetric difference between the sets of selected and annotated features (Nguyen & Martínez, 2020). A particularly common metric involves summing the attribution values over known important features, often with normalization or weighting adjustments[i].

For **visual and natural language domains**, the well-known *Pointing Game* checks whether the most highly attributed features lie within predefined ground-truth regions such as bounding boxes or key tokens[j]. Closely related is *Weakly Supervised Localization*, where the explanans is compared directly to segmentation masks or bounding boxes, typically using Intersection-over-Union (IoU)[k]. In multi-label or multi-object scenarios, further alignment can be tested by verifying whether the explanation focuses on the input features associated with the predicted label (Du et al., 2019).

For **CEs**, evaluation typically involves test datasets that explicitly annotate the presence or absence of each concept, enabling precise assessment of concept identification accuracy (Kim et al., 2018; Lucieri et al., 2020; Asokan et al., 2022).

For **NLEs**, comparisons to reference justifications rely on general natural language processing measures such as BLEU or ROUGE[l], as well as similarity measures specifically designed for NLEs (Park et al., 2018; Du et al., 2022)

Finally, to ensure that the evaluation isolates explanation quality from predictive accuracy, Fan et al. (2022) recommend to limit evaluation to those samples for which the model's prediction is both correct and sufficiently confident.

[a] Kim et al. (2018); Arras et al. (2019); Yang & Kim (2019); Tritscher et al. (2020); Faber et al. (2021)

[b] Chen et al. (2017); Oramas et al. (2017); Ross et al. (2017); Chen et al. (2018b); Yang et al. (2018b); Ying et al. (2019); Luo et al. (2020); Faber et al. (2021); Kim et al. (2021); Tjoa & Guan (2022); Wilming et al. (2022); Zhou et al. (2022)

[c] Cortez & Embrechts (2013); Jia et al. (2019; 2020); Liu et al. (2021a); Amoukou et al. (2022)

[d] Oramas et al. (2017); Ross et al. (2017); Kim et al. (2018); Yang et al. (2018b); Ying et al. (2019); Lucieri et al. (2020); Luo et al. (2020); Faber et al. (2021); Kim et al. (2021); Yalcin et al. (2021); Khakzar et al. (2022); Tjoa & Guan (2022); Agarwal et al. (2023); Hesse et al. (2023); Miró-Nicolau et al. (2023)

[e] Lin et al. (2019; 2021); Fan et al. (2022); Sun et al. (2023); Ya et al. (2023)

[f] Cortez & Embrechts (2013); Chen et al. (2017; 2018b); Ismail et al. (2019); Jia et al. (2019); Amiri et al. (2020); Ismail et al. (2020); Jia et al. (2020); Nguyen & Martínez (2020); Tritscher et al. (2020); Liu et al. (2021a); Agarwal et al. (2022b); Amoukou et al. (2022)

[g] Bohle et al. (2021); Shah et al. (2021); Arias-Duart et al. (2022); Rao et al. (2022; 2024)

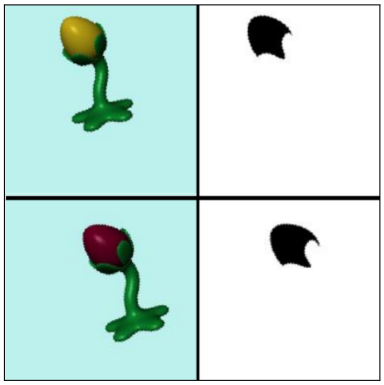
[h] Selvaraju et al. (2017); Kanehira et al. (2019); Wang & Nvasconcelos (2019); Cheng et al. (2020); Saifullah et al. (2024)

[i]Lapuschkin et al. (2016); Yang & Kim (2019); Kohlbrenner et al. (2020); Nam et al. (2020); Rio-Torto et al. (2020); Wang et al. (2020a); Xu et al. (2020a); Bohle et al. (2021); Kim et al. (2021); Arias-Duart et al. (2022); Arras et al. (2022); Zhou et al. (2022)

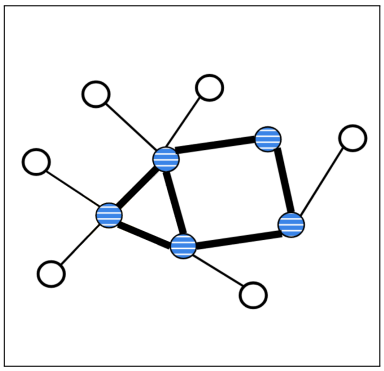
[j]Bargal et al. (2018); Poerner et al. (2018); Zhang et al. (2018a); Fong et al. (2019); Sydorova et al. (2019); Taghanaki et al. (2019); Takeishi (2019); Schulz et al. (2020); Barnett et al. (2021); Arras et al. (2022); Theiner et al. (2022)

[k]Simonyan et al. (2013); Cao et al. (2015); Zhou et al. (2016); Fong & Vedaldi (2017); Selvaraju et al. (2017); Wickramanayake et al. (2019); Bany Muhammad & Yeasin (2021); Nguyen et al. (2021); Fan et al. (2022)

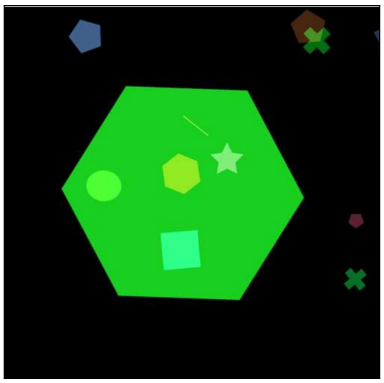
[l]Camburu et al. (2018); Chuang et al. (2018); Liu et al. (2018a); Wu & Mooney (2018); Chen et al. (2019d); Rajani et al. (2019); Wickramanayake et al. (2019); Li et al. (2020a); Sun et al. (2020); Jang & Lukasiewicz (2021); Wiegreffe & Marasović (2021); Ribeiro et al. (2023); Atanasova (2024b)



(a) The an8Flower dataset (Oramas et al., 2017).
Left: Input images of flowers, with differing petals. Right: Ground-Truth explanans of relevant region.



(b) The ShapeGGen dataset (Agarwal et al., 2023).
The graph's label depends on relevant structures or motifs (e.g. house-shape) that serve as ground-truth explanans.



(c) The SCDB dataset (Lucieri et al., 2020).
The label is defined by the combination of small objects inside the are of the larger one and serve as ground-truth concepts.

Figure 8: Examples of synthetic datasets containing ground-truth explanantia of relevant features.

Metric V.41: White Box Model Check

Desiderata: Fidelity
Explanation Type: FA, (ExE, NLE) **References:** (12)
Ribeiro et al. (2016); Antwarg et al. (2019); Dhurandhar et al. (2019); Jia et al. (2019); Zhou et al. (2019); Crabbe et al. (2020); Jia et al. (2020); Guidotti (2021); Velmurugan et al. (2021a); Dai et al. (2022); Brandt et al. (2023); Carmichael & Scheirer (2023)

In white-box models, the internal logic is fully accessible and interpretable, providing a ground-truth rationale against which generated explanantia can be directly evaluated. Typical white-box models used for comparison include linear regressors (Crabbe et al., 2020; Dai et al., 2022), feature-additive models (Carmichael & Scheirer, 2023), small neural networks with manually set parameters (Antwarg et al., 2019; Brandt et al., 2023), or symbolic models such as decision trees (Ribeiro et al., 2016; Dhurandhar et al., 2019).
Given that the model's reasoning is known, explanation methods can be assessed by comparing their output against this ground-truth explanans. This can be done using general similarity metrics (Guidotti, 2021; Jia et al., 2019) or using error metrics like MSE (Crabbe et al., 2020; Dai et al., 2022; Brandt et al., 2023).
Further, if the white-box model relies only on a constrained feature subset, one can also measure explanation fidelity through accuracy, precision, or recall between explanans and the truly influential features (Ribeiro et al., 2016; Zhou et al., 2019; Jia et al., 2020; Velmurugan et al., 2021a).
Although originally proposed for FAs, this approach may be extended to ExEs and NLEs, as the white-box model enables verification of whether the provided explanantia are consistent with the

known reasoning. While user inspection is straightforward, adapting this check into an automatic, functionality-grounded evaluation remains challenging but potentially feasible.

C Excluded Metrics

We excluded a total of four metrics from our framework. While initially identified during our literature review, we later dismissed them after closer inspection. Some do not evaluate the explanation itself, but rather the model or downstream tasks, while others rely on assumptions that are overly narrow or reward properties we consider misleading. For transparency, we briefly describe each and illustrate our reasons for exclusion below.

- **Alignment**: Etmann et al. (2019) assess plausibility by measuring how well a saliency map aligns with the original input, using correlation as a proxy. However, we argue that input reconstruction does not necessarily enhance interpretability, as simply returning the input or an edge map offers little explanatory value. Furthermore, the metric mixes explanation with input similarity, rewarding explanation that may lack selectivity or meaningful abstraction. Therefore, we exclude this metric due to its flawed underlying assumption.
- **Explanatory Power** Arras et al. (2017) evaluate the model–explanation pair by computing k-nearest-neighbor accuracy on document vectors constructed from word-attribution explanation. While the score reflects how semantically useful these explanations are for related tasks, it measures downstream utility rather than explanation quality. As such, it evaluates a combination of model and explanation performance, not the fidelity (or other desideratum) of the explanation itself. Further, it is not applicable beyond text domain or in scenarios without downstream tasks.
- **Feature Diversity**: Smyth & Keane (2022) calculate the fraction of different features altered across all counterfactuals in a dataset, where low diversity implies the same features are consistently changed. However, this appears to be a purely statistical property of the explanation, and it remains unclear why higher or lower diversity should indicate better explanation quality. Without a clear link to any desideratum, its practical value as an evaluation metric is doubtful.
- **Mutual Information**: Le et al. (2020) measure how independent the features changed in counterfactuals are, with the goal of encouraging maximal expressiveness. It uses Symmetrical Uncertainty (Press, 2007) to quantify pairwise dependencies among changed features. However, enforcing independence between altered features can easily lead to implausible counterfactuals; especially when high correlations reflect natural dependencies. Breaking these can result in unrealistic or invalid inputs.

References

Carlo Abrate and Francesco Bonchi. Counterfactual graphs for explainable classification of brain networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2495–2504, 2021.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pp. 50–59. PMLR, 2018.

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations. *arXiv preprint arXiv:2203.06877*, 2022a.

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, 2022b.

Chirag Agarwal, Marinka Zitnik, and Himabindu Lakkaraju. Probing gnn explainers: A rigorous theoretical and empirical analysis of gnn explanation methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 8969–8996. PMLR, 2022c.

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, 2023.

Nourah Alangari, Mohamed El Bachir Menai, Hassan Mathkour, and Ibrahim Almosallam. Exploring evaluation methods for interpretable machine learning: A survey. *Information*, 14(8):469, 2023a.

Nourah Alangari, Mohamed El Bachir Menai, Hassan Mathkour, and Ibrahim Almosallam. Intrinsically interpretable gaussian mixture model. *Information*, 14(3):164, 2023b.

Emanuele Albini, Antonio Rago, Pietro Baroni, Francesca Toni, et al. Relation-based counterfactual explanations for bayesian network classifiers. In *IJCAI*, pp. 451–457, 2020.

Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1054–1070, 2022.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805, 2023.

José M Alonso, Luis Magdalena, and Serge Guillaume. Hilk: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism. *International Journal of Intelligent Systems*, 23(7):761–794, 2008.

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

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks, 2018b. URL https://arxiv.org/abs/1806.07538.

Shideh Shams Amiri, Rosina O Weber, Prateek Goel, Owen Brooks, Archer Gandley, Brian Kitchell, and Aaron Zehm. Data representing ground-truth explanations to evaluate xai methods. *arXiv preprint arXiv:2011.09892*, 2020.

Salim I Amoukou, Tangi Salaün, and Nicolas Brunel. Accurate shapley values for explaining tree-based models. In *International conference on artificial intelligence and statistics*, pp. 2448–2465. PMLR, 2022.

Elvio Amparore, Alan Perotti, and Paolo Bajardi. To trust or not to trust an explanation: using leaf to evaluate local linear xai methods. *PeerJ Computer Science*, 7:e479, 2021.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International Conference on Machine Learning*, pp. 272–281. PMLR, 2019.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2020.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 382–398. Springer, 2016.

Robert Andrews, Joachim Diederich, and Alan B Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems*, 8(6):373–389, 1995.

Raghuram Mandyam Annasamy and Katia Sycara. Towards better interpretability in deep q-networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4561–4569, 2019.

Liat Antwarg, Ronnie Mindlin Miller, Bracha Shapira, and Lior Rokach. Explaining anomalies detected by autoencoders using shap. *arXiv preprint arXiv:1903.02407*, 2019.

Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Giménez-Ábalos. Focus! rating xai methods and finding biases. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8. IEEE, 2022.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. " what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142, 2017.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. *arXiv preprint arXiv:1904.11829*, 2019.

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

Osvaldo Arreche, Tanish R Guntur, Jack W Roberts, and Mustafa Abdallah. E-xai: Evaluating black-box explainable ai frameworks for network intrusion detection. *IEEE Access*, 2024.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

André Artelt and Barbara Hammer. Convex density constraints for computing plausible counterfactual explanations. In *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I 29*, pp. 353–365. Springer, 2020.

André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating robustness of counterfactual explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 01–09. IEEE, 2021.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

Ashish Ramayee Asokan, Nidarshan Kumar, Anirudh V Ragam, and SS Shylaja. Interpretability for multimodal emotion recognition using concept activation vectors. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 01–08. IEEE, 2022.

Pepa Atanasova. A diagnostic study of explainability techniques for text classification. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pp. 155–187. Springer, 2024a.

Pepa Atanasova. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pp. 83–103. Springer, 2024b.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10445–10453, 2022.

M Gethsiyal Augasta and Thangairulappan Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, 35:131–150, 2012.

Md Abdul Awal and Chanchal K Roy. Evaluatexai: A framework to evaluate the reliability and consistency of rule-based xai techniques for software analytics tasks. *Journal of Systems and Software*, 217:112159, 2024.

Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612*, 2020.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. *Advances in Neural Information Processing Systems*, 34:5644–5655, 2021.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Naman Bansal, Chirag Agarwal, and Anh Nguyen. Sam: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 8673–8683, 2020.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Visual explanations for deep convolutional neural networks. *SN Computer Science*, 2(1):47, 2021.

Nahla Barakat, Andrew P Bradley, and Mohamed Nabil H Barakat. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, 14(4): 1114–1120, 2010.

Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for rnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1440–1449, 2018.

Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. Interpretable mammographic image classification using case-based reasoning and deep learning. *arXiv preprint arXiv:2107.05605*, 2021.

Cher Bass, Mariana da Silva, Carole Sudre, Petru-Daniel Tudosiu, Stephen Smith, and Emma Robinson. Icam: Interpretable classification via disentangled representations and feature attribution mapping. *Advances in Neural Information Processing Systems*, 33:7697–7709, 2020.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*, 2017.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160*, 2019.

Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8309–8319, 2018.

Betül Bayrak and Kerstin Bach. Pertcf: A perturbation-based counterfactual generation approach. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 174–187. Springer, 2023a.

Betül Bayrak and Kerstin Bach. A twin xcbr system using supportive and contrastive explanations. In *ICCBR 2023 Workshop Proceedings*. CEUR Workshop Proceedings, 2023b.

Betül Bayrak and Kerstin Bach. Evaluation of instance-based explanations: An in-depth analysis of counterfactual evaluation metrics, challenges, and the ceval toolkit. *IEEE Access*, 2024.

Punam Bedi, Anjali Thukral, and Shivani Dhiman. Explainable ai in disease diagnosis. In *Explainable AI in Health Informatics*, pp. 87–111. Springer, 2024.

Mohamed Karim Belaid, Eyke Hüllermeier, Maximilian Rabus, and Ralf Krestel. Do we need another explainable ai method? toward unifying post-hoc xai evaluation methods into an interactive and multidimensional benchmark. *arXiv preprint arXiv:2207.14160*, 2022.

Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. Archived at: https://web.archive.org/web/20250210101243/https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Alexander Binder, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16143–16152, 2023.

Jessica Y Bo, Pan Hao, and Brian Y Lim. Incremental xai: Memorable understanding of ai with incremental explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023.

Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10029–10038, 2021.

Philine Lou Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M-C Höhne. Finding the right xai method—a guide for the evaluation and ranking of explainable ai methods in climate science. *Artificial Intelligence for the Earth Systems*, 3(3):e230074, 2024.

Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pp. 1014–1023. PMLR, 2020.

Mohammed Brahimi, Saïd Mahmoudi, Kamel Boukhalfa, and Abdelouhab Moussaoui. Deep interpretable architecture for plant diseases classification. In *2019 signal processing: Algorithms, architectures, arrangements, and applications (SPA)*, pp. 111–116. IEEE, 2019.

Rafaël Brandt, Daan Raatjens, and Georgi Gaydadjiev. Precise benchmarking of explainable ai attribution methods. *arXiv preprint arXiv:2308.03161*, 2023.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Lukas Brunke, Prateek Agrawal, and Nikhil George. Evaluating input perturbation methods for interpreting cnns and saliency map comparison. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 120–134. Springer, 2020.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pp. 454–464, 2020.

Kirill Bykov, Marina M-C Höhne, Adelaida Creosteanu, Klaus-Robert Müller, Frederick Klauschen, Shinichi Nakajima, and Marius Kloft. Explaining bayesian neural networks. *arXiv preprint arXiv:2108.10346*, 2021.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. Can i trust the explainer? verifying post-hoc explanatory methods. *arXiv preprint arXiv:1910.02065*, 2019.

Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2956–2964, 2015.

Zachariah Carmichael and Walter J Scheirer. How well do feature-additive explainers explain feature-additive predictors? *arXiv preprint arXiv:2310.18496*, 2023.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and characterizing human rationales. *arXiv preprint arXiv:2010.04736*, 2020.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Ahmad Chaddad, Yan Hu, Yihang Wu, Binbin Wen, and Reem Kateb. Generalizable and explainable deep learning for medical image computing: An overview. *Current Opinion in Biomedical Engineering*, pp. 100567, 2024.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pp. 1383–1391. PMLR, 2020.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Chao Chen, Yifei Liu, Xi Zhang, and Sihong Xie. Scalable explanation of inferences on large graphs. In *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 982–987. IEEE, 2019a.

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 world wide web conference*, pp. 1583–1592, 2018a.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint arXiv:2004.02015*, 2020.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. Can rationalization improve robustness? *arXiv preprint arXiv:2204.11790*, 2022.

Jianbo Chen, Mitchell Stern, Martin J Wainwright, and Michael I Jordan. Kernel feature selection via conditional covariance minimization. *Advances in neural information processing systems*, 30, 2017.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*, pp. 883–892. PMLR, 2018b.

Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, and Somesh Jha. Robust attribution regularization. *Advances in Neural Information Processing Systems*, 32, 2019b.

Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 765–774, 2019c.

Yinkai Chen, Rui Zhang, Xin Qiu, Xin Li, and Yuxin Deng. Explaining gbdt by probabilistic finite-state automata. In *Proceedings of the 2021 10th International Conference on Computing and Pattern Recognition*, pp. 328–333, 2021.

Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. Co-attentive multi-task learning for explainable recommendation. In *IJCAI*, volume 2019, pp. 2137–2143, 2019d.

Weiyu Cheng, Yanyan Shen, Linpeng Huang, and Yanmin Zhu. Incorporating interpretability into latent factor models via fast influence analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 885–893, 2019.

Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12925–12935, 2020.

Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

Michael Chromik and Martin Schuessler. A taxonomy for human subject evaluation of black-box explanations in xai. *Exss-atec@ iui*, 1:1–7, 2020.

Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang, and Jian Pei. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1244–1253, 2018.

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 975–983, 2018.

Tobias Clement, Nils Kemmerzell, Mohamed Abdelaal, and Michael Amberg. Xair: A systematic metareview of explainable ai (xai) aligned to the software development process. *Machine Learning and Knowledge Extraction*, 5(1):78–108, 2023.

Zicun Cong, Lingyang Chu, Lanjun Wang, Xia Hu, and Jian Pei. Exact and consistent interpretation of piecewise linear models hidden behind apis: A closed form solution. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 613–624. IEEE, 2020.

Loredana Coroama and Adrian Groza. Evaluation metrics in explainable artificial intelligence (xai). In *International conference on advanced research in technologies, information, innovation and sustainability*, pp. 401–413. Springer, 2022.

Paulo Cortez and Mark J Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17, 2013.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Jonathan Crabbe, Yao Zhang, William Zame, and Mihaela van der Schaar. Learning outside the black-box: The pursuit of interpretable models. *Advances in neural information processing systems*, 33:17838–17849, 2020.

Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8, 1995.

Robin Cugny, Julien Aligon, Max Chevalier, Geoffrey Roman Jimenez, and Olivier Teste. Autoxai: A framework to automatically select the most adapted xai solution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 315–324, 2022.

Yiming Cui, Ting Liu, Wanxiang Che, Zhigang Chen, and Shijin Wang. Expmrc: explainability evaluation for machine reading comprehension. *Heliyon*, 8(4), 2022.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

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.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International conference on parallel problem solving from nature*, pp. 448–469. Springer, 2020.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

Devleena Das, Sonia Chernova, and Been Kim. State2explanation: Concept-based explanations to benefit agent learning and user understanding. *Advances in Neural Information Processing Systems*, 36:67156–67182, 2023.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. In *International Conference on Machine Learning*, pp. 4794–4815. PMLR, 2022.

Luiz Felipe De Camargo, Diego Roberto Colombo Dias, and José Remo Ferreira Brega. Implementation of explainable artificial intelligence: Case study on the assessment of movements to support neuromotor rehabilitation. In *International Conference on Computational Science and Its Applications*, pp. 564–580. Springer, 2023.

Mithsen De Silva, Prabhash Kumarasinghe, Kasun De Zoysa, and Chamath Keppitiyagama. Reidentifying asian elephants from ear images using a cascade of convolutional neural networks and explaining with gradcam. *SN Computer Science*, 3(3):192, 2022.

Eoin Delaney, Derek Greene, and Mark T Keane. Instance-based counterfactual explanations for time series classification. In *International conference on case-based reasoning*, pp. 32–47. Springer, 2021.

Houtao Deng. Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, 7(4):277–287, 2019.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. Model agnostic contrastive explanations for structured data. *arXiv preprint arXiv:1906.00117*, 2019.

Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *ECAI 2020*, pp. 2473–2480. IOS Press, 2020.

Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, and Ahmed M Ali. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, 615: 238–292, 2022.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

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

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*, 2022.

Mengnan Du, Ninghao Liu, Fan Yang, Shuiwang Ji, and Xia Hu. On attribution of recurrent neural network predictions via additive decomposition. In *The world wide web conference*, pp. 383–393, 2019.

Jiarui Duan, Haoling Li, Haofei Zhang, Hao Jiang, Mengqi Xue, Li Sun, Mingli Song, and Jie Song. On the evaluation consistency of attribution-based explanations. *arXiv preprint arXiv:2407.19471*, 2024.

Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pp. 1–6, 2019.

Radwa El Shawi, Youssef Sherif, and Sherif Sakr. Towards automated concept-based decision treeexplanations for cnns. In *EDBT*, pp. 379–384, 2021.

Ghada Elkhawaga, Omar Elzeki, Mervat Abuelkheir, and Manfred Reichert. Evaluating explainable artificial intelligence methods based on feature elimination: A functionality-grounded approach. *Electronics*, 12(7): 1670, 2023.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

European Union High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Technical report, European Commission, 2019. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Benjamin P Evans, Bing Xue, and Mengjie Zhang. What's inside the black-box? a genetic programming method for interpreting complex machine learning models. In *Proceedings of the genetic and evolutionary computation conference*, pp. 1012–1020, 2019.

Lukas Faber, Amin K. Moghaddam, and Roger Wattenhofer. When comparing to ground truth is wrong: On evaluating gnn explanation methods. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 332–341, 2021.

Ming Fan, Wenying Wei, Xiaofei Xie, Yang Liu, Xiaohong Guan, and Ting Liu. Can we trust your explanations? sanity checks for interpreters in android malware analysis. *IEEE Transactions on Information Forensics and Security*, 16:838–853, 2020.

Ming Fan, Jiali Wei, Wuxia Jin, Zhou Xu, Wenying Wei, and Ting Liu. One step further: evaluating interpreters using metamorphic testing. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 327–339, 2022.

Nawshad Farruque, Randy Goebel, Osmar R Zaïane, and Sudhakar Sivapalan. Explainable zero-shot modelling of clinical depression symptoms from text. In *2021 20th IEEE international conference on machine learning and applications (ICMLA)*, pp. 1472–1477. IEEE, 2021.

Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. A study on the interpretability of neural retrieval models using deepshap. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pp. 1005–1008, 2019.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2950–2958, 2019.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Nazanin Fouladgar, Marjan Alirezaie, and Kary Främling. Metrics and evaluations of time series explanations: An application in affect computing. *IEEE Access*, 10:23995–24009, 2022.

Matija Franklin and David Lagnado. Human-ai interaction paradigm for evaluating explainable artificial intelligence. In *International Conference on Human-Computer Interaction*, pp. 404–411. Springer, 2022.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

Thorben Funke, Megha Khosla, Mandeep Rathee, and Avishek Anand. Zorro: Valid, sparse, and stable explanations in graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 35(8): 8687–8698, 2022.

Francesco Fusco, Michalis Vlachos, Vasileios Vasileiadis, Kathrin Wardatzky, and Johannes Schneider. Reconet: An interpretable neural architecture for recommender systems. In *IJCAI*, pp. 2343–2349, 2019.

Yingqiang Ge, Shuchang Liu, Zelong Li, Shuyuan Xu, Shijie Geng, Yunqi Li, Juntao Tan, Fei Sun, and Yongfeng Zhang. Counterfactual evaluation for explainable ai. *arXiv preprint arXiv:2109.01962*, 2021.

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.

Felipe Giuste, Wenqi Shi, Yuanda Zhu, Tarun Naren, Monica Isgut, Ying Sha, Li Tong, Mitali Gupte, and May D Wang. Explainable artificial intelligence methods in combating pandemics: A systematic review. *IEEE Reviews in Biomedical Engineering*, 16:5–21, 2022.

Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. A survey of explainable ai techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*, pp. 1–37, 2024.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Mara Graziani, Thomas Lompech, Henning Müller, and Vincent Andrearczyk. Evaluation and comparison of cnn visual explanations for histopathology. In *Proceedings of the AAAI Conference on Artificial Intelligence Workshops (XAI-AAAI-21), Virtual Event*, pp. 8–9, 2021.

Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of cnns via contrastive backpropagation. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 119–134. Springer, 2019.

Riccardo Guidotti. Evaluating local explanation methods on ground truth. *Artificial Intelligence*, 291: 103428, 2021.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6): 14–23, 2019.

David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40 (2):44–58, 2019.

Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. *arXiv preprint arXiv:2012.15781*, 2020.

Tian Guo, Tao Lin, and Nino Antulov-Fantulin. Exploring interpretable lstm neural networks over multi-variable data. In *International conference on machine learning*, pp. 2494–2504. PMLR, 2019.

Wenbo Guo, Sui Huang, Yunzhe Tao, Xinyu Xing, and Lin Lin. Explaining deep learning models–a bayesian non-parametric approach. *Advances in neural information processing systems*, 31, 2018a.

Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. Lemna: Explaining deep learning based security applications. In *proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pp. 364–379, 2018b.

Isha Hameed, Samuel Sharpe, Daniel Barcklow, Justin Au-Yeung, Sahil Verma, Jocelyn Huang, Brian Barr, and C Bayan Bruss. Based-xai: Breaking ablation studies down for explainable artificial intelligence. *arXiv preprint arXiv:2207.05566*, 2022.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*, 2020.

Satoshi Hara and Kohei Hayashi. Making tree ensembles interpretable: A bayesian model selection approach. In *International conference on artificial intelligence and statistics*, pp. 77–85. PMLR, 2018.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*, 2020.

Julian Hatwell, Mohamed Medhat Gaber, and R Muhammad Atif Azad. Ada-whips: explaining adaboost classification with applications in the health sciences. *BMC Medical Informatics and Decision Making*, 20: 1–25, 2020.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina Höhne. Sanity checks revisited: An exploration to repair the model parameter randomisation test. *arXiv preprint arXiv:2401.06465*, 2024.

Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact recruiter's decision. *IEEE Transactions on Affective Computing*, 14(2):969–985, 2021.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in neural information processing systems*, 32, 2019.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3981–3991, 2023.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations. *KI-Künstliche Intelligenz*, 34(2):193–198, 2020.

Milo Honegger. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054*, 2018.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Md Tanzib Hosain, Jamin Rahman Jim, MF Mridha, and Md Mohsin Kabir. Explainable ai approaches in deep learning: Advancements, applications and challenges. *Computers and Electrical Engineering*, 117: 109246, 2024.

Janet Hui-wen Hsiao, Hilary Hei Ting Ngai, Luyu Qiu, Yi Yang, and Caleb Chen Cao. Roadmap of designing cognitive metrics for explainable artificial intelligence (xai). *arXiv preprint arXiv:2108.01737*, 2021.

Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. Evaluations and methods for explanation through robustness analysis. *arXiv preprint arXiv:2006.00442*, 2020.

Wei Huang, Xingyu Zhao, Gaojie Jin, and Xiaowei Huang. Safari: Versatile and efficient evaluations for robustness of interpretability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1988–1998, 2023a.

Zexi Huang, Mert Kosan, Sourav Medya, Sayan Ranu, and Ambuj Singh. Global counterfactual explainer for graph neural networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 141–149, 2023b.

Frederik Hvilshøj, Alexandros Iosifidis, and Ira Assent. On quantitative evaluations of counterfactuals. *arXiv preprint arXiv:2111.00177*, 2021.

Adrian Höhl, Ivica Obadic, Miguel-Ángel Fernández-Torres, Hiba Najjar, Dario Augusto Borges Oliveira, Zeynep Akata, Andreas Dengel, and Xiao Xiang Zhu. Opening the black box: A systematic review on explainable artificial intelligence in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 12 (4):261–304, 2024. doi: 10.1109/MGRS.2024.3467001.

Rami Ibrahim and M Omair Shafiq. Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys*, 55(10):1–37, 2023.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1511–1519, 2019.

Aya Abdelsalam Ismail, Mohamed Gunady, Luiz Pessoa, Hector Corrada Bravo, and Soheil Feizi. Input-cell attention reduces vanishing saliency of recurrent neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems*, 33: 6441–6452, 2020.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Myeongjun Jang and Thomas Lukasiewicz. Are training resources insufficient? predict first then explain! *arXiv preprint arXiv:2110.02056*, 2021.

Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 805–815, 2021.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1459–1467. PMLR, 2021.

Kishlay Jha, Yaqing Wang, Guangxu Xun, and Aidong Zhang. Interpretable word embeddings for medical domain. In *2018 IEEE international conference on data mining (ICDM)*, pp. 1061–1066. IEEE, 2018.

Yunzhe Jia, James Bailey, Kotagiri Ramamohanarao, Christopher Leckie, and Michael E Houle. Improving the quality of explanations with local embedding perturbations. In *Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & Data Mining*, pp. 875–884, 2019.

Yunzhe Jia, James Bailey, Kotagiri Ramamohanarao, Christopher Leckie, and Xingjun Ma. Exploiting patterns to explain individual predictions. *Knowledge and Information Systems*, 62:927–950, 2020.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14901–14909, 2023.

Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical image analysis*, 84:102684, 2023.

Ulf Johansson, Rikard König, and Lars Niklasson. The truth is in there-rule extraction from opaque models using genetic programming. In *FLAIRS*, pp. 658–663, 2004.

Indu Joshi, Riya Kothari, Ayush Utkarsh, Vinod K Kurmi, Antitza Dantcheva, Sumantra Dutta Roy, and Prem Kumar Kalra. Explainable fingerprint roi segmentation using monte carlo dropout. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 60–69, 2021.

Hyungsik Jung and Youngrock Oh. Towards better explanations of class activation mapping. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1336–1344, 2021.

Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. Evaluation metrics for xai: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pp. 000111–000124. IEEE, 2023.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *IJCAI*, pp. 2855–2862, 2020.

Atsushi Kanehira and Tatsuya Harada. Learning to explain with complemental examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8603–8611, 2019.

Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting counterfactuality in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8594–8602, 2019.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4948–4957, 2019.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics*, pp. 895–905. PMLR, 2020.

Isak Karlsson, Jonathan Rebane, Panagiotis Papapetrou, and Aristides Gionis. Explainable time series tweaking via irreversible and reversible temporal transformations. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 207–216. IEEE, 2018.

Ashkan Khakzar, Pedram Khorsandi, Rozhin Nobahari, and Nassir Navab. Do explanations explain? model knows best. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10244–10253, 2022.

Aaishwarya Khalane, Rikesh Makwana, Talal Shaikh, and Abrar Ullah. Evaluating significant features in context-aware multimodal emotion recognition with xai methods. *Expert Systems*, pp. e13403, 2023.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Sanity simulations for saliency methods. *arXiv preprint arXiv:2105.06506*, 2021.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 267–280, 2019.

Lukas Klein, Carsten T Lüth, Udo Schlegel, Till J Bungert, Mennatallah El-Assady, and Paul F Jäger. Navigating the maze of explainable ai: A systematic approach to evaluating methods and metrics. *arXiv preprint arXiv:2409.16756*, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

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

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.

Sanjay Krishnan and Eugene Wu. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2Nd workshop on human-in-the-loop data analytics*, pp. 1–6, 2017.

I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pp. 5491–5500. PMLR, 2020.

Aditya Kuppa and Nhien-An Le-Khac. Black box attacks on explainable artificial intelligence (xai) methods in cyber security. In *2020 International Joint Conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1675–1684, 2016.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 131–138, 2019.

Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *International conference on machine learning*, pp. 5628–5638. PMLR, 2020.

Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021.

Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2912–2920, 2016.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv:1806.07498*, 2018.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Issues with post-hoc counterfactual explanations: a discussion. *arXiv preprint arXiv:1906.04774*, 2019a.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294*, 2019b.

Phuong Quynh Le, Meike Nauta, Shreyasi Pathak Van Bach Nguyen, Shreyasi Pathak, Jörg Schlötterer, and Christin Seifert. Benchmarking explainable ai-a survey on available toolkits and open challenges. In *IJCAI*, pp. 6665–6673, 2023.

Thai Le, Suhang Wang, and Dongwon Lee. Grace: Generating concise and informative contrastive sample to explain neural network model's prediction. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 238–248, 2020.

Anastasia-M Leventi-Peetz and Kai Weber. Rashomon effect and consistency in explainable artificial intelligence (xai). In *Proceedings of the Future Technologies Conference*, pp. 796–808. Springer, 2022.

Lei Li, Yongfeng Zhang, and Li Chen. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 755–764, 2020a.

Runze Li, Tomaso Fontanini, Andrea Prati, and Bir Bhanu. Face synthesis with a focus on facial attributes translation using attention mechanisms. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1):76–90, 2022.

Xiao-Hui Li, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, et al. A survey of data-driven and knowledge-aware explainable ai. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):29–49, 2020b.

Xiao-Hui Li, Yuhan Shi, Haoyang Li, Wei Bai, Yuanwei Song, Caleb Chen Cao, and Lei Chen. Quantitative evaluations on saliency methods: An experimental study. *arXiv preprint arXiv:2012.15616*, 2020c.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *NeurIPS*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yi-Shan Lin, Wen-Chuan Lee, and Z Berkay Celik. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1027–1035, 2021.

Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint arXiv:1910.07387*, 2019.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable nlp: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196*, 2018a.

Ninghao Liu, Xiao Huang, Jundong Li, and Xia Hu. On interpretation of network embedding via taxonomy induction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1812–1820, 2018b.

Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8642–8651, 2020.

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

Yifei Liu, Chao Chen, Yazheng Liu, Xi Zhang, and Sihong Xie. Multi-objective explanations of gnn predictions. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 409–418. IEEE, 2021b.

Helena Löfström, Karl Hammar, and Ulf Johansson. A meta survey of quality evaluation criteria in explanation methods. In *International Conference on Advanced Information Systems Engineering*, pp. 55–63. Springer, 2022.

Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19):9423, 2022.

Adriano Lucieri, Muhammad Naseer Bajwa, Andreas Dengel, and Sheraz Ahmed. Explaining ai-based decision support systems using concept localization maps. In *International Conference on Neural Information Processing*, pp. 185–193. Springer, 2020.

Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine*, 215:106620, 2022.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33: 19620–19631, 2020.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. Leveraging latent features for local explanations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1139–1149, 2021.

Jing Ma, Ruocheng Guo, Saumitra Mishra, Aidong Zhang, and Jundong Li. Clear: Generative counterfactual explanations on graphs. *Advances in neural information processing systems*, 35:25895–25907, 2022.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Sarthak Malik, Puneet Kumar, and Balasubramanian Raman. Towards interpretable facial emotion recognition. In *Proceedings of the twelfth Indian conference on computer vision, graphics and image processing*, pp. 1–9, 2021.

Vincent Margot and George Luta. A new method to compare the interpretability of rule-based algorithms. *Ai*, 2(4):621–635, 2021.

Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113:103655, 2021.

Joao Marques-Silva, Thomas Gerspacher, Martin Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. *Advances in Neural Information Processing Systems*, 33:20590–20600, 2020.

David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4942–4950, 2018.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14867–14875, 2021.

Dominique Mercier, Jwalin Bhatt, Andreas Dengel, and Sheraz Ahmed. Time to focus: A comprehensive benchmark using time series attribution methods. *arXiv preprint arXiv:2202.03759*, 2022.

Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–7. IEEE, 2019.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40, 2022.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pp. 1–66, 2022.

Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. A novel approach to generate datasets with xai ground truth to evaluate image models. *arXiv preprint arXiv:2302.05624*, 2023.

Saumitra Mishra, Emmanouil Benetos, Bob LT Sturm, and Simon Dixon. Reliable local explanations for machine listening. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding human knowledge into deep neural network via attention map. *arXiv preprint arXiv:1905.03540*, 2019.

Elhassan Mohamed, Konstantinos Sirlantzis, and Gareth Howells. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. *Displays*, 73:102239, 2022.

Yusuf Abas Mohamed, Bee Ee Khoo, Mohd Shahrimie Mohd Asaari, Mohd Ezane Aziz, and Fattah Rahiman Ghazali. Decoding the black box: Explainable ai (xai) for cancer diagnosis, prognosis, and treatment planning-a state-of-the art systematic review. *International Journal of Medical Informatics*, pp. 105689, 2024.

Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4): 1–45, 2021.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 417–431. Springer, 2020.

Milad Moradi and Matthias Samwald. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, 165:113941, 2021.

Sofia Morandini, Federico Fraboni, Gabriele Puzzo, Davide Giusino, Lucia Volpi, Hannah Brendel, Enzo Balatti, Marco De Angelis, Andrea De Cesarei, and Luca Pietrantoni. Examining the nexus between explainability of ai systems and user's trust: A preliminary scoping review. *Proceedings http://ceur-ws.org ISSN*, 1613:0073, 2023.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

Sascha Mücke and Lukas Pfahler. Check mate: A sanity check for trustworthy ai. In *LWDA*, pp. 91–103, 2022.

Ricardo Müller, Marco Schreyer, Timur Sattarov, and Damian Borth. Reshape: explaining accounting anomalies in financial statement audits by enhancing shapley additive explanations. In *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 174–182, 2022.

Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 2501–2508, 2020.

Detlef D Nauck. Measuring interpretability in rule-based classification systems. In *The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ'03.*, volume 1, pp. 196–201. IEEE, 2003.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Sidra Naveed, Gunnar Stevens, and Dean Robin-Kern. An overview of the empirical evaluation of explainable ai (xai): A comprehensive guideline for user-centered evaluation in xai. *Applied Sciences*, 14(23):11288, 2024.

Hossein Nematzadeh, José García-Nieto, Ismael Navas-Delgado, and José F Aldana-Montes. Ensemble-based genetic algorithm explainer with automized image segmentation: A case study on melanoma detection dataset. *Computers in Biology and Medicine*, 155:106613, 2023.

Hillary Ngai and Frank Rudzicz. Doctor xavier: Explainable diagnosis on physician-patient dialogues and xai evaluation. *arXiv preprint arXiv:2204.10178*, 2022.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020.

Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1069–1078. Association for Computational Linguistics, 2018.

Hung Truong Thanh Nguyen, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pp. 1–6, 2021.

Thu Trang Nguyen, Thach Le Nguyen, and Georgiana Ifrim. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*, pp. 77–94. Springer, 2020.

Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International conference on machine learning*, pp. 3809–3818. PMLR, 2018.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018.

Jose Oramas, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. *arXiv preprint arXiv:1712.06302*, 2017.

Evandro S Ortigossa, Thales Gonçalves, and Luis Gustavo Nonato. Explainable artificial intelligence (xai)–from theory to methods and applications. *IEEE Access*, 2024.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, 2021.

Sebastian Palacio, Adriano Lucieri, Mohsin Munir, Sheraz Ahmed, Jörn Hees, and Andreas Dengel. Xai handbook: towards a unified framework for explainable ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3766–3775, 2021.

Deng Pan, Xiangrui Li, Xin Li, and Dongxiao Zhu. Explainable recommendation via interpretable feature mapping and evaluation of explainability. *arXiv preprint arXiv:2007.06133*, 2020a.

Menghai Pan, Weixiao Huang, Yanhua Li, Xun Zhou, and Jun Luo. xgail: Explainable generative adversarial imitation learning for explainable human decision analysis. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1334–1343, 2020b.

Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 629–639, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8779–8788, 2018.

Serin Park and Christian Wallraven. Comparing facial expression recognition in humans and machines: Using cam, gradcam, and extremal perturbation. In *Asian Conference on Pattern Recognition*, pp. 403–416. Springer, 2021.

Cristiano Patrício, João C Neves, and Luís F Teixeira. Explainable deep learning methods in medical image classification: A survey. *ACM Computing Surveys*, 56(4):1–41, 2023.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of the web conference 2020*, pp. 3126–3132, 2020.

Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: a python library to benchmark algorithmic recourse and counterfactual explanation algorithms. *arXiv preprint arXiv:2108.00783*, 2021.

Marek Pawlicki, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'Antonio, Rafał Kozik, and Michał Choraś. Evaluating the necessity of the multiple metrics for assessing explainable ai: A critical examination. *Neurocomputing*, 602:128282, 2024.

Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27 (7):1173–1185, 2020.

Georgina Peake and Jun Wang. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2060–2069, 2018.

Tejaswini Pedapati, Avinash Balakrishnan, Karthikeyan Shanmugam, and Amit Dhurandhar. Learning global transparent models consistent with local contrastive explanations. *Advances in neural information processing systems*, 33:3592–3602, 2020.

Uwe Peters and Mary Carman. Cultural bias in explainable ai research: A systematic analysis. *Journal of Artificial Intelligence Research*, 79:971–1000, 2024.

V Petsiuk. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Juan D Pinto and Luc Paquette. Towards a unified framework for evaluating explanations. *arXiv preprint arXiv:2405.14016*, 2024.

Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31, 2018.

Gregory Plumb, Jonathan Terhorst, Sriram Sankararaman, and Ameet Talwalkar. Explaining groups of points in low-dimensional representations. In *International Conference on Machine Learning*, pp. 7762–7771. PMLR, 2020.

Nina Poerner, Benjamin Roth, and Hinrich Schütze. Evaluating neural network explanation methods using hybrid documents and morphological agreement. *arXiv preprint arXiv:1801.06422*, 2018.

Mirko Polato and Fabio Aiolli. Interpretable preference learning: A game theoretic framework for large margin on-line feature and rule learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4723–4730, 2019.

Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10772–10781, 2019.

Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2299–2304, 2021.

Chanathip Pornprasit, Chakkrit Tantithamthavorn, Jirayus Jiarpakdee, Michael Fu, and Patanamon Thongtanunam. Pyexplainer: Explaining the predictions of just-in-time defect models. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 407–418. IEEE, 2021.

William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*, 2019.

Nikaash Puri, Sukriti Verma, Piyush Gupta, Dhruv Kayastha, Shripad Deshmukh, Balaji Krishnamurthy, and Sameer Singh. Explain your move: Understanding agent actions using specific and relevant feature attribution. *arXiv preprint arXiv:1912.12191*, 2019.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019.

Dilini Rajapaksha, Christoph Bergmeir, and Wray Buntine. Lormika: Local rule-based model interpretability with k-optimal associations. *Information Sciences*, 540:221–241, 2020.

Gopi Krishnan Rajbahadur, Shaowei Wang, Gustavo A Oliva, Yasutaka Kamei, and Ahmed E Hassan. The impact of feature importance methods on the interpretation of defect classifiers. *IEEE Transactions on Software Engineering*, 48(7):2245–2261, 2021.

Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. *Advances in Data Analysis and Classification*, 14:801–819, 2020.

Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10223–10232, 2022.

Sukrut Rao, Moritz Böhle, and Bernt Schiele. Better understanding differences in attribution methods via systematic evaluations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396, 2022.

Maria H Rasmussen, Diana S Christensen, and Jan H Jensen. Do machines dream of atoms? crippen's logp as a quantitative molecular benchmark for explainable ai heatmaps. *SciPost Chemistry*, 2(1):002, 2023.

Peyman Rasouli and Ingrid Chieh Yu. Analyzing and improving the robustness of tabular classifiers using counterfactual explanations. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1286–1293. IEEE, 2021.

Jay Raval, Pronaya Bhattacharya, Nilesh Kumar Jadav, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, Mitwalli Elmorsy, Amr Tolba, and Maria Simona Raboaca. Raksha: A trusted explainable lstm model to classify fraud patterns on credit card transactions. *Mathematics*, 11(8):1901, 2023.

Atul Rawal, James McCoy, Danda B Rawat, Brian M Sadler, and Robert St Amant. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 3(6):852–866, 2021.

Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, 33:12187–12198, 2020.

Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8839–8848, 2020.

Danilo Ribeiro, Shen Wang, Xiaofei Ma, Henry Zhu, Rui Dong, Deguang Kong, Juliette Burger, Anjelica Ramos, William Wang, Zhiheng Huang, et al. Street: A multi-task structured reasoning and explanation benchmark. *arXiv preprint arXiv:2302.06729*, 2023.

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.

Mireia Ribera and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *Proceedings of the Human-Centered AI Workshop (HCXAI)*, volume 2327 of *CEUR Workshop Proceedings*, 2019. URL http://ceur-ws.org/Vol-2327/. Accessed: 2024-08-07.

Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. *arXiv preprint arXiv:2003.08747*, 2020.

Isabel Rio-Torto, Kelwin Fernandes, and Luís F Teixeira. Understanding the decisions of cnns: An in-model approach. *Pattern Recognition Letters*, 133:373–380, 2020.

Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pp. 159–175, 2018.

Nikolaos Rodis, Christos Sardianos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Iraklis Varlamis, and Georgios Th Papadopoulos. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. *IEEE Access*, 2024.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.

Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th international conference on autonomous agents and multiagent systems*, pp. 45–50, 2021.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Raif Rustamov and James Klosowski. Interpretable graph-based semi-supervised learning via flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Mirka Saarela and Vili Podgorelec. Recent applications of explainable ai (xai): A systematic literature review. *Applied Sciences*, 14(19):8884, 2024.

Saifullah Saifullah, Dominique Mercier, Adriano Lucieri, Andreas Dengel, and Sheraz Ahmed. The privacy-explainability trade-off: unraveling the impacts of differential privacy and federated learning on attribution methods. *Frontiers in Artificial Intelligence*, 7:1236947, 2024.

Ahmed Salih, Ilaria Boscolo Galazzo, Federica Cruciani, Lorenza Brusini, and Petia Radeva. Investigating explainable artificial intelligence for mri-based classification of dementia: a new stability criterion for explainable methods. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 4003–4007. IEEE, 2022.

Ahmed M Salih, Ilaria Boscolo Galazzo, Polyxeni Gkontra, Elisa Rauseo, Aaron Mark Lee, Karim Lekadir, Petia Radeva, Steffen E Petersen, and Gloria Menegaz. A review of evaluation approaches for explainable ai with applications in cardiology. *Artificial Intelligence Review*, 57(9):240, 2024a.

Ahmed M Salih, Ilaria Boscolo Galazzo, Zahra Raisi-Estabragh, Steffen E Petersen, Gloria Menegaz, and Petia Radeva. Characterizing the contribution of dependent features in xai methods. *IEEE Journal of Biomedical and Health Informatics*, 2024b.

Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, 2019.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wesley Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating attribution for graph neural networks. *Advances in neural information processing systems*, 33:5898–5910, 2020.

Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. Explainable ai for healthcare 5.0: opportunities and challenges. *IEEE Access*, 10:84486–84517, 2022.

Dominik Schiller, Tobias Huber, Michael Dietz, and Elisabeth André. Relevance-based data masking: a model-agnostic transfer learning approach for facial expression recognition. *Frontiers in Computer Science*, 2:6, 2020.

Udo Schlegel and Daniel A Keim. A deep dive into perturbations as evaluation technique for time series xai. In *World Conference on Explainable Artificial Intelligence*, pp. 165–180. Springer, 2023.

Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A Keim. Towards a rigorous evaluation of xai methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4197–4201. IEEE, 2019.

Udo Schlegel, Daniela Oelke, Daniel A Keim, and Mennatallah El-Assady. An empirical study of explainable ai techniques on deep learning models for time series tasks. *arXiv preprint arXiv:2012.04344*, 2020.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020.

Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pp. 1–59, 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.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems*, 34:2046–2059, 2021.

Farhad Shakerin and Gopal Gupta. Induction of non-monotonic logic programs to explain boosted tree models using lime. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3052–3059, 2019.

Lloyd S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, pp. 307–317. Princeton University Press, Princeton, NJ, 1953.

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 166–172, 2020.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 395–405, 2019.

Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, 39(7):1390–1404, 2023.

Ilija Šimić, Vedran Sabol, and Eduardo Veas. Perturbation effect: A metric to counter misleading validation of feature attribution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1798–1807, 2022.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018.

Jaspreet Singh, Zhenye Wang, Megha Khosla, and Avishek Anand. Valid explanations for learning to rank models. *arXiv preprint arXiv:2004.13972*, 2020.

Jaspreet Singh, Megha Khosla, Wang Zhenye, and Avishek Anand. Extracting per query valid explanations for blackbox learning-to-rank models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 203–210, 2021.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483*, 2019.

Xuelin Situ, Ingrid Zukerman, Cecile Paris, Sameen Maruf, and Gholamreza Haffari. Learning to explain: Generating stable explanations fast. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5340–5355, 2021.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *International conference on machine learning*, pp. 9046–9057. PMLR, 2020.

Dylan Slack, Sorelle A Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501*, 2019.

Barry Smyth and Mark T Keane. A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International Conference on Case-Based Reasoning*, pp. 18–32. Springer, 2022.

Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 56–67, 2020.

Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. A survey on methods and metrics for the assessment of explainability under the proposed ai act. In *Legal Knowledge and Information Systems*, pp. 235–242. IOS Press, 2021.

Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 2239–2250, 2022.

Jerzy Stefanowski and Daniel Vanderpooten. Induction of decision rules in classification and discovery-oriented perspectives. *International Journal of Intelligent Systems*, 16(1):13–27, 2001.

Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martin Pereira-Fariña. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8. IEEE, 2020.

Alexander Stevens and Johannes De Smedt. Explainability in process outcome prediction: Guidelines to obtain interpretable and faithful models. *European Journal of Operational Research*, 317(2):317–329, 2024.

Marilyn Strathern. 'improving ratings': audit in the british university system. *European review*, 5(3): 305–321, 1997.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. Obtaining faithful interpretations from compositional neural networks. *arXiv preprint arXiv:2005.00724*, 2020.

Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2020–2029, 2019.

Jimin Sun, Wenqi Shi, Felipe O Giuste, Yog S Vaghani, Lingzi Tang, and May D Wang. Improving explainable ai with patch perturbation-based evaluation pipeline: a covid-19 x-ray image analysis case study. *Scientific reports*, 13(1):19488, 2023.

Peijie Sun, Le Wu, Kun Zhang, Yanjie Fu, Richang Hong, and Meng Wang. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. In *Proceedings of The Web Conference 2020*, pp. 837–847, 2020.

Yi Sun, Abel Valente, Sijia Liu, and Dakuo Wang. Preserve, promote, or attack? gnn explanation via topology perturbation. *arXiv preprint arXiv:2103.13944*, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Alona Sydorova, Nina Poerner, and Benjamin Roth. Interpretable question answering on knowledge bases and text. *arXiv preprint arXiv:1906.10924*, 2019.

Saeid Asgari Taghanaki, Mohammad Havaei, Tess Berthier, Francis Dutil, Lisa Di Jorio, Ghassan Hamarneh, and Yoshua Bengio. Infomask: Masked variational latent representation to localize chest disease. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pp. 739–747. Springer, 2019.

Naoya Takeishi. Shapley values of reconstruction errors of pca for explaining anomaly detection. In *2019 international conference on data mining workshops (icdmw)*, pp. 793–798. IEEE, 2019.

Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM web conference 2022*, pp. 1018–1027, 2022.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 303–310, 2018.

Sule Tekkesinoglu and Sara Pudas. Explaining graph convolutional network predictions for clinicians—an explainable ai approach to alzheimer's disease classification. *Frontiers in Artificial Intelligence*, 6:1334613, 2024.

Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 750–760, 2022.

Andreas Theissler, Francesco Spinnato, Udo Schlegel, and Riccardo Guidotti. Explainable ai for time series classification: a review, taxonomy and research directions. *Ieee Access*, 10:100700–100724, 2022.

Erico Tjoa and Cuntai Guan. Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset. *IEEE Transactions on Artificial Intelligence*, 4(4):858–870, 2022.

Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 465–474, 2017.

Nicholay Topin and Manuela Veloso. Generation of policy-level explanations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2514–2521, 2019.

Julian Tritscher, Markus Ring, Daniel Schlr, Lena Hettinger, and Andreas Hotho. Evaluation of post-hoc xai approaches through synthetic tabular data. In *Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings*, pp. 422–430. Springer, 2020.

Julian Tritscher, Maximilian Wolf, Andreas Hotho, and Daniel Schlör. Evaluating feature relevance xai in network intrusion detection. In *World Conference on Explainable Artificial Intelligence*, pp. 483–497. Springer, 2023.

Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Perception of image features in post-mortem iris recognition: Humans vs machines. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–8. IEEE, 2019.

U.S. National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical report, U.S. Department of Commerce, 2023. URL http://dx.doi.org/10.6028/NIST.AI.100-1.

Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial intelligence*, 291:103404, 2021.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 650–665. Springer, 2021.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Manjunatha Veerappa, Mathias Anneken, Nadia Burkart, and Marco F Huber. Validation of xai explanations for multivariate time series classification in the maritime domain. *Journal of Computational Science*, 58: 101539, 2022.

Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Developing a fidelity evaluation approach for interpretable machine learning. *arXiv preprint arXiv:2106.08492*, 2021a.

Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Evaluating fidelity of explainable methods for predictive process analytics. In *International conference on advanced information systems engineering*, pp. 64–72. Springer, 2021b.

Manisha Verma and Debasis Ganguly. Lirme: locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1281–1284, 2019.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12):1–42, 2024.

Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa De Oliveira, and David Martens. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, 25(2):315–335, 2022.

Tom Viering, Ziqi Wang, Marco Loog, and Elmar Eisemann. How to manipulate cnns to make them lie: the gradcam case. *arXiv preprint arXiv:1907.10901*, 2019.

Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

Minh N Vu, Truc D Nguyen, NhatHai Phan, Ralucca Gera, and My T Thai. c-eval: A unified metric to evaluate feature-based explanations via perturbation. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 927–937. IEEE, 2021.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9097–9107, 2019.

Eric Wang, Pasha Khosravi, and Guy Van den Broeck. Probabilistic sufficient explanations. *arXiv preprint arXiv:2105.10118*, 2021.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020a.

Pei Wang and Nuno Nvasconcelos. Deliberative explanations: visualizing network insecurities. *Advances in Neural Information Processing Systems*, 32, 2019.

Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8981–8990, 2020.

Tong Wang. Multi-value rule sets for interpretable classification with feature-efficient representations. *Advances in neural information processing systems*, 31, 2018.

Tong Wang. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*, pp. 6505–6514. PMLR, 2019.

Wei Wang, Christopher Lesner, Alexander Ran, Marko Rukonic, Jason Xue, and Eric Shiu. Using small business banking data for explainable credit risk scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13396–13401, 2020b.

Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. A reinforcement learning framework for explainable recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 587–596. IEEE, 2018a.

Yulong Wang, Hang Su, Bo Zhang, and Xiaolin Hu. Interpret neural networks by identifying critical data routing paths. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8906–8914, 2018b.

Alexander Warnecke, Daniel Arp, Christian Wressnegger, and Konrad Rieck. Evaluating explanation methods for deep learning in security. In *2020 IEEE european symposium on security and privacy (EuroS&P)*, pp. 158–174. IEEE, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. Flex: Faithful linguistic explanations for neural net based model decisions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 2539–2546, 2019.

Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint arXiv:2102.12060*, 2021.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

Rick Wilming, Céline Budding, Klaus-Robert Müller, and Stefan Haufe. Scrutinizing xai using linear ground-truth data with suppressor variables. *Machine learning*, 111(5):1903–1923, 2022.

Huijun Wu, Chen Wang, Jie Yin, Kai Lu, and Liming Zhu. Sharing deep neural network models with interpretation. In *Proceedings of the 2018 World Wide Web Conference*, pp. 177–186, 2018a.

Jialin Wu and Raymond J Mooney. Faithful multimodal explanation for visual question answering. *arXiv preprint arXiv:1809.02805*, 2018.

Mike Wu, Michael Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018b.

Mike Wu, Sonali Parbhoo, Michael Hughes, Ryan Kindle, Leo Celi, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Regional tree regularization for interpretability in deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6413–6421, 2020.

Qianqian Xie, Prayag Tiwari, Deepak Gupta, Jimin Huang, and Min Peng. Neural variational sparse topic model for sparse explainable text representation. *Information Processing & Management*, 58(5):102614, 2021.

Wenjia Xu, Jiuniu Wang, Yang Wang, Guangluan Xu, Daoyu Lin, Wei Dai, and Yirong Wu. Where is the model looking at?–concentrate and explain the network attention. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):506–516, 2020a.

Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9523–9532, 2020b.

Biao Xua and Guanci Yang. Interpretability research of deep learning: A literature survey. *Information Fusion*, pp. 102721, 2024.

Mengxi Ya, Yiming Li, Tao Dai, Bin Wang, Yong Jiang, and Shu-Tao Xia. Towards faithful xai evaluation via generalization-limited backdoor watermark. In *The Twelfth International Conference on Learning Representations*, 2023.

Orcun Yalcin, Xiuyi Fan, and Siyuan Liu. Evaluating the correctness of explainable ai algorithms for classification. *arXiv preprint arXiv:2105.09740*, 2021.

Fan Yang, Ninghao Liu, Suhang Wang, and Xia Hu. Towards interpretation of recommender systems with sorted explanation paths. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 667–676. IEEE, 2018a.

Fan Yang, Mengnan Du, and Xia Hu. Evaluating explanation without ground truth in interpretable machine learning. *arXiv preprint arXiv:1907.06831*, 2019.

Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019.

Yinchong Yang, Volker Tresp, Marius Wunderle, and Peter A Fasching. Explaining therapy predictions with layer-wise relevance propagation in neural networks. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 152–162. IEEE, 2018b.

Chih-Kuan Yeh and Pradeep Ravikumar. Objective criteria for explanations of machine learning models. *Applied AI Letters*, 2(4):e57, 2021.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the sensitivity and stability of model interpretations in nlp. *arXiv preprint arXiv:2104.08782*, 2021.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Jaemin Yoo and Lee Sael. Edit: Interpreting ensemble models via compact soft decision trees. In *2019 IEEE international conference on data mining (ICDM)*, pp. 1438–1443. IEEE, 2019.

Hao Yuan, Yongjun Chen, Xia Hu, and Shuiwang Ji. Interpreting deep models for text analysis via optimization and regularization methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5717–5724, 2019.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5782–5799, 2022.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013. URL https://arxiv.org/abs/1311.2901.

Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8660–8669, 2019.

Hao Zhang, Jiayi Chen, Haotian Xue, and Quanshi Zhang. Towards a unified evaluation of explanation methods without ground truth. *arXiv preprint arXiv:1911.09017*, 2019a.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018a.

Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018b.

Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6261–6270, 2019b.

Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. Interpreting neural network judgments via minimal, stable, and symbolic corrections. *Advances in neural information processing systems*, 31, 2018c.

Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In *29th {USENIX} security symposium ({USENIX} security 20)*, 2020.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

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.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9623–9633, 2022.

Zihan Zhou, Mingxuan Sun, and Jianhua Chen. A model-agnostic approach for explaining the predictions on clustered data. In *2019 IEEE international conference on data mining (ICDM)*, pp. 1528–1533. IEEE, 2019.

Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. Deepred–rule extraction from deep neural networks. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19*, pp. 457–473. Springer, 2016.

Lin Zou, Han Leong Goh, Charlene Jin Yee Liew, Jessica Lishan Quah, Gary Tianyu Gu, Jun Jie Chew, Mukundaram Prem Kumar, Christine Gia Lee Ang, and Andy Wee An Ta. Ensemble image explainable ai (xai) algorithm for severe community-acquired pneumonia and covid-19 respiratory infections. *IEEE Transactions on Artificial Intelligence*, 4(2):242–254, 2022.