A Novel Approach for Benchmarking Local Binary Classification XAI Methods Using Synthetic Ground Truth Datasets

Karim Moustafa[1]

[1]*The Artificial Intelligence and Cognitive Load Research Lab, The Centre of Explainable Artificial Intelligence, Technological University Dublin, Dublin, Republic of Ireland*

Abstract

Evaluating Explainable AI (XAI) methods, particularly local feature attributions on tabular data, is hindered by a lack of standardised benchmarks and objective metrics. This research introduces a novel technique to address this gap. We propose a methodology for generating synthetic ground-truth datasets with known feature-target relationships, enabling objective evaluation. The proposed solution includes the Explainability Fidelity Assessor (XFA) to quantify interpretation faithfulness (completeness, conciseness, sensitivity) and the Optimum Explainability Score (OPS) to assess objective explanation quality (simplicity, completeness). Preliminary results using LIME and Anchor demonstrate the ability to compare XAI methods objectively. This work contributes to the creation of XAI groundtruth datasets and the development of novel evaluation metrics (XFA, OPS), aiming to enhance the reliability and comparability of local XAI methods.

1. Context and Motivation

Machine learning (ML) models are increasingly deployed in critical domains, yet complex models often function as black boxes, hindering understanding of their decision-making processes. Explainable Artificial Intelligence (XAI) aims to address this by providing insights into model behaviour [1]. Understanding *how* a model makes a prediction is crucial for debugging, building trust, ensuring fairness, promoting accountability, and enabling user interaction [2, 3]. For instance, in healthcare, knowing which patient features drive a high-risk prediction is vital for clinical decision-making [4].

Despite the proliferation of XAI methods, evaluating their effectiveness remains a significant challenge [3]. A key problem is the lack of standardised benchmark datasets with known ground truth, particularly for evaluating *local* explanations (explanations for individual predictions). Without such datasets, comparing different XAI techniques objectively is difficult. Furthermore, there's no universally accepted definition or framework for assessing explanation quality, leading to inconsistency and concerns about generalisability [5, 6]. The field often conflates inherent model *interpretability* with post-hoc *explainability* as in [7, 8]. It sometimes relies on subjective human evaluations [9], which, however, are costly, difficult to scale, and prone to bias [2].

This research tackles these challenges by proposing a novel technique for evaluating local feature attributions of XAI methods applied to tabular, cross-sectional data for binary classification. The core contributions are: (1) The introduction of the first synthetic ground-truth datasets specifically designed for local XAI evaluation, providing an objective basis for assessment. (2) A novel evaluation framework comprising the Explainability Fidelity Assessor (XFA) for measuring faithfulness to the ground truth, and the Optimum Explainability Score (OPS) for assessing objective explanation characteristics like simplicity and completeness. By providing standardised datasets and objective metrics, this work aims to improve the reliability, comparability, and trustworthiness of XAI methods.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

d15127673@mytudublin.ie (K. Moustafa)
https://www.linkedin.com/in/karimmoustafa/ (K. Moustafa)
0000-0002-8636-9946 (K. Moustafa)

2. Key Related Work

The need for transparency in complex ML models has driven the development of numerous XAI techniques [10]. These methods can be categorised along several dimensions:

- **Applicability:** Model-specific (tailored to specific architectures, e.g., activation analysis in CNNs [11]) versus model-agnostic (treat model as black-box, e.g., LIME [12], SHAP [13], Anchor [14]). This work focuses on model-agnostic methods due to their flexibility.
- **Scope:** Global (overall model behaviour) versus local (individual predictions). This research targets local explanations explicitly.
- **Stage:** Intrinsic (inherently simple models like decision trees [15]) versus post-hoc (techniques applied after training complex models). Our focus is on evaluating post-hoc methods for black-box models.
- **Output Form:** Explanations can be feature importance scores, rules, counterfactuals, visualisations [16]. This work primarily targets methods that produce feature importance scores or similar quantifiable outputs.

Evaluating XAI methods is crucial but challenging [3]. Existing evaluation approaches often fall into categories proposed by Doshi-Velez and Kim [3]:

- **Human-grounded:** Uses user studies to assess understanding, trust, satisfaction [2]. Pro: Direct human insight. Con: Subjective, costly, hard to scale/generalise.
- **Application-grounded:** Evaluates XAI impact on task performance in a specific domain [17]. Pro: Realistic context. Con: It is hard to isolate XAI impact; findings may not generalise.
- **Function-grounded:** Uses proxy metrics like fidelity, stability, robustness [18, 19]. Pro: Objective, automatable. Con: Metrics may not reflect true human interpretability.

Recent efforts have included developing benchmark datasets, primarily for NLP, based on human annotations [20], which inherit subjectivity and lack generalizability to other domains, such as tabular data. Significant gaps remain in XAI evaluation [21, 22]. There is:

- **Lack of Consensus and Generic Metrics:** No agreed-upon definition of ”good” explanation or standard metrics applicable across different XAI methods [5].
- **Confusion between Interpretability and Explainability:** Over-reliance on subjective human evaluation obscures the need for objective assessment of explanation fidelity.
- **Limited Comparability:** Difficulty in comparing different XAI algorithms (macro-level) or tuning parameters within an algorithm (micro-level) due to a lack of standard benchmarks.
- **Lack of Holistic Local Evaluation:** No unified way to aggregate performance across different local instances and evaluation aspects.

This research directly addresses the need for objective, generic metrics and standardised ground-truth benchmarks for local feature attributions of XAI methods on tabular data.

3. Research Questions and Objectives

This research aims to develop an objective benchmarking technique for evaluating local feature attribution explanations methods for binary classification on tabular cross-sectional data. The central research question is:

Can an XAI evaluation benchmarking technique—based on (1) a synthetic ground-truth dataset, (2) explainability fidelity assessment, and (3) optimum local explainability assessment—effectively evaluate the feature attribution performance of local XAI methods for binary classification?

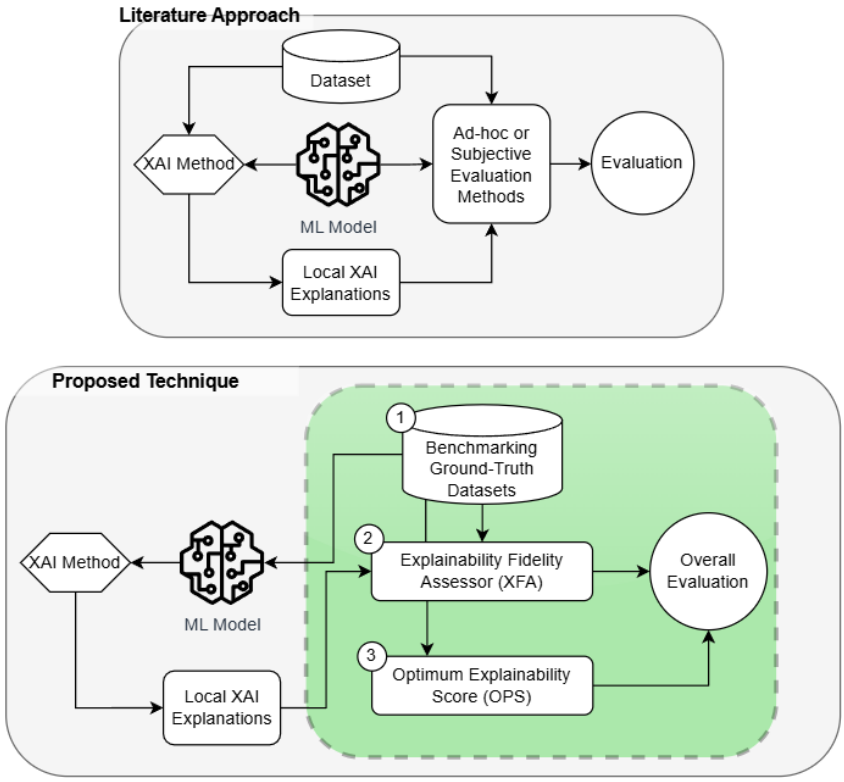

Figure 1: Proposed Technique Components vs. Literature Approaches.

4. Research Approach, Methods, and Rationale

To address the identified gaps and research questions, we propose a novel evaluation technique focused on objective assessment using synthetic ground truth. The proposed solution workflow, contrasted with typical literature approaches, is shown in Figure 1. It separates the generation of ground truth from the evaluation of fidelity (XFA) and objective explanation characteristics (OPS).

4.1. Synthetic Ground-truth Datasets Creation

Rationale: The lack of objective ground truth is a major impediment to XAI evaluation. Synthetic datasets allow us to precisely control the underlying feature-target relationships. *Method:* We generate datasets where the binary target variable is determined by predefined 'Relationship Rules' (RRs) applied to specific subsets of features. These rules encompass linear and non-linear regression-based functions, as well as distance-based clustering structures. Table 2 shows the summary of the RRs used in dataset generation.

The process involves:

1. **Dataset Pool Creation:** Generate datasets covering combinations of 2-12 RRs. Each dataset (10k instances, 48 features) uses multivariate normal distribution features. Instances are stratified, and labels are generated per stratum using the assigned RR. Crucially, instance-level metadata (the ground truth specifying which features and rules generated the label) is stored.
2. **Dataset Selection:** Train a standard ML model (e.g., RandomForest) on each dataset of the 4082 datasets created in the previous step, as a (80/20 split). Select 11 datasets representing the 0th to 100th percentiles of model performance (e.g., AUC), ensuring a range of complexities (Table 1). These selected datasets form the benchmark suite.

This provides the first benchmark suite with known instance-level ground truth for local XAI evaluation on tabular data.

Table 1
Selected Datasets Complexity Scale (Examples).

RRs Combination Used	Complexity Scale
(1, 3)	1 (Lowest Complexity)
(1, 5, 6, 11, 12)	3
(1, 2, 3, 5, 6, 8, 9, 10, 11, 12)	7
(5, 6, 7, 9, 10)	11 (Highest Complexity)
... (11 datasets selected in total)	

Table 2
Summary of Relationship Rules (RRs) for Synthetic Dataset Generation

RR ID	Type	Basis / Key Characteristic	Description Summary	# Feat.
1	Regression	Linear, $y = f(x_1)$	Monotonic single variable	1
2	Regression	Linear, $y = f(x_1, x_2)$	Monotonic two variables	2
3	Regression	Non-linear, $y = f(x_1^2, x_2, x_3)$	Weighted non-linear combination	3
4	Regression	Hastie (Spheres), $y = f(\sum x_i^2)$	Concentric spheres	2
5	Regression	Hastie (Spheres), $y = f(\sum x_i^2)$	Concentric spheres	5
6	Regression	Hastie (Chi-squared), $y = f(\sum x_j^2)$	Threshold on sum of squares	10
7	Regression	Friedman 1 (Non-linear)	Based on sin, $(x - c)^2$, linear terms	5
8	Regression	Friedman 2 (Non-linear)	Simulates circuit impedance	4
9	Regression	Friedman 3 (Non-linear)	Simulates circuit phase shift (arctan)	4
10	Distance	MADELON (Hypercube)	Clusters at hypercube vertices	4
11	Distance	Gaussian Blobs	Two isotropic Gaussian clusters	6
12	Distance	Moons	Two interleaving half-circles	2

4.2. Explainability Fidelity Assessor (XFA)

Rationale: To objectively measure how faithfully an XAI method's explanation reflects the true underlying reasons (ground truth) for a prediction. *Method:* XFA compares the features identified as important by an XAI method (for a given instance) against the ground truth features from the dataset metadata for that instance (Figure 2). It computes three instance-level metrics, which are then averaged across the test set:

- **Completeness (IF-Recall):** Percentage of ground truth features correctly identified by the XAI method. (High recall = finds most true features).
- **Conciseness (1 - IF-FPR):** Inverse of the False Positive Rate. Percentage of features identified by XAI that are *actually* in the ground truth. (High conciseness = identifies fewer incorrect features).
- **Sensitivity (IF-Sen):** Agreement between the rank order of feature importance given by XAI and the rank order in the ground truth (if applicable, e.g., based on coefficients).

The final **XFA Score** aggregates these metrics, weighted by dataset complexity, providing a single score (0-100) for overall fidelity as shown in (Eq. 1).

$$XFA\_Score = \sum_{i}^{n} avg([1 - IF\text{-}FPR_i], IF\text{-}Recall_i, IF\text{-}Sen_i) \\ * DatasetComplexity_i * ComplexityScaler \tag{1}$$

4.3. Optimum Explainability Score (OPS)

Rationale: To evaluate objective properties of the explanation itself, namely its simplicity and its coverage (completeness of application). Aim is to find the "sweet spot" between being informative and being concise. *Method:* OPS considers two factors:

- **Simplicity (C_{xai}):** Average number of features (or explanation elements) included in the local explanations across the dataset. Lower is simpler.

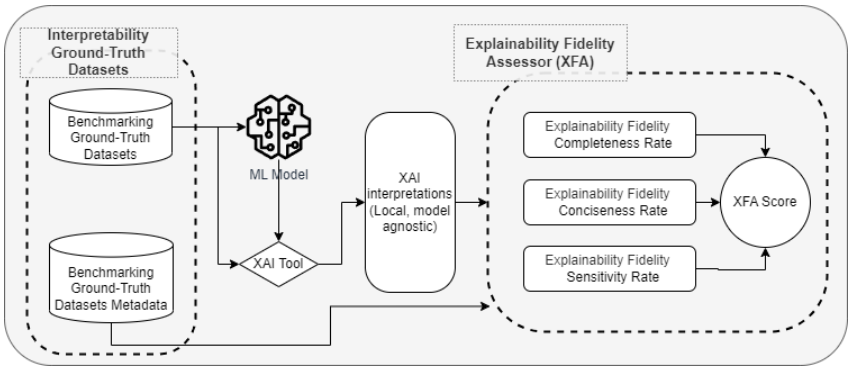

Figure 2: Components of the Explainability Fidelity Assessor (XFA).

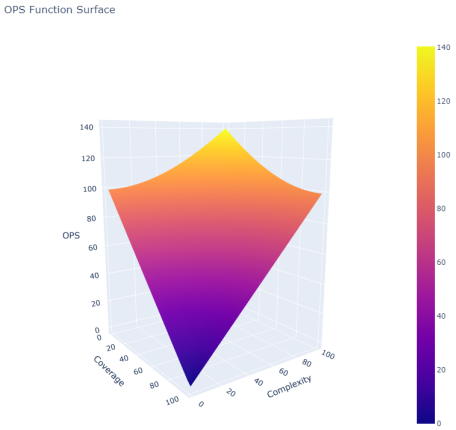

Figure 3: Surface plot of the Optimum Explainability Score (OPS). Lower (blue) is better.

- **Completeness (V_{xai}):** Percentage of instances for which the XAI method successfully generated an explanation meeting a predefined quality/confidence threshold (e.g., Anchor's precision threshold). Higher is better.

The **OPS score** (Eq. 2) measures the Euclidean distance from the ideal point (0 complexity, 100

$$OPS = \sqrt{(100 - V_{xai})^2 + C_{xai}^2} - 1 \tag{2}$$

4.4. Overall Score and Validation

The **Overall Score** combines XFA and OPS (normalised) via a weighted average, providing a holistic evaluation as shown in (Eq.3), with user-adjustable weights (W_1, W_2) enabling prioritisation of specific evaluation aspects.

$$OverallScore = \frac{1}{2}\left(W_1\ XFA + \left(W_2\ e^{\frac{-OPS}{100}} * 100\right)\right) \tag{3}$$

Validation Rationale: To ensure the proposed metrics (XFA, OPS) behave as expected and are sensitive to changes in explanation quality. *Validation Method:*

- **XFA Validation:** Systematically introduce controlled errors (missing features, extra features, incorrect ranking) into the ground truth metadata at varying levels (0-100%). Assess if the XFA score and its components show a monotonic decrease with increasing error levels and a strong association (correlation) with the error level.

- **OPS Validation:** Generate explanation sets with varying controlled levels of complexity (C_{xai}) and coverage (V_{xai}). Assess if the OPS score correctly ranks these explanation sets and correlates with the deviation from the ideal (0 complexity, 100% coverage).

5. Preliminary Results and Contributions to Date

Progress has been achieved in developing the proposed XAI evaluation benchmarking framework:

- A comprehensive literature review identified critical gaps in current XAI evaluation approaches.
- The methodology for generating synthetic ground truth datasets was designed and implemented, resulting in a benchmark suite of 11 datasets with varying, controllable complexity levels and known instance-level ground truth.
- The Explainability Fidelity Assessor (XFA) component was designed and developed. Its core properties (e.g., monotonicity, association with controlled error) were validated using the protocol outlined in Section 4, confirming its suitability for objective fidelity assessment.
- The Optimum Explainability Score (OPS) component for evaluating objective explanation characteristics (simplicity, coverage) was designed, and its underlying mathematical properties (convexity) were established.

Demonstrating Framework Utility: XFA Use Case To illustrate the practical application and capabilities of our framework, we conducted an initial case study using the XFA component and the synthetic datasets. This involved comparing two widely recognised local, model-agnostic explanation methods: LIME [12] and Anchor [14]. The process involved generating explanations from both methods for instances in our benchmark datasets and evaluating their fidelity using XFA against the known ground truth (see Figure 4 for the workflow).

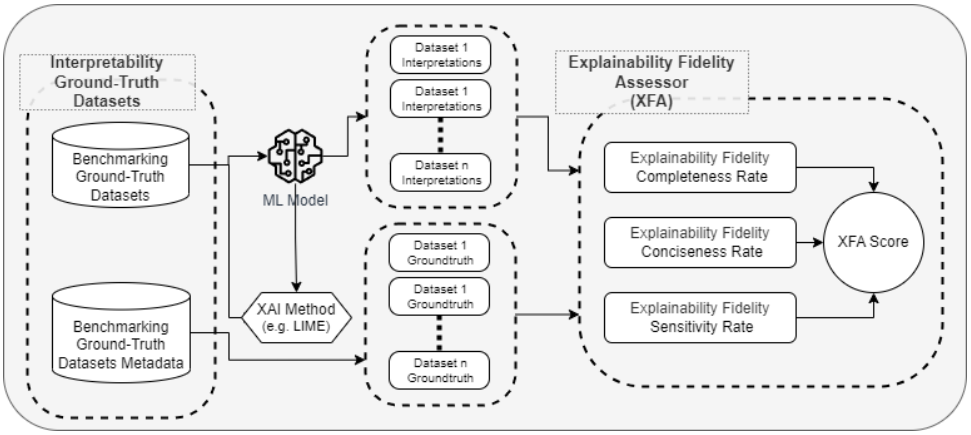

Figure 4: XFA use case process for comparing LIME and Anchor.

Qualitative Insights from the Use Case: This preliminary application demonstrated that the XFA framework successfully enables:

- **Objective Quantitative Comparison:** The framework provided distinct quantitative fidelity scores for LIME and Anchor based on metrics assessing completeness (finding true features), conciseness (avoiding false positives), and sensitivity (ranking features correctly) versus ground truth.
- **Analysis of Fidelity Profiles:** The comparison highlighted nuanced differences in how these methods perform. For example, it allows for the analysis of trade-offs, such as potential differences in achieving high recall versus maintaining high conciseness in the generated explanations.
- **Assessment of Complexity Impact:** The experiments using datasets of varying complexity indicated that the fidelity of explanations can be influenced by the complexity of the underlying data-generating process, and the framework allows for observing these effects.

- **Highlighting Areas for Improvement:** The overall assessment suggested that while current methods provide valuable insights, there is measurable variance in their fidelity and potential scope for enhancing the faithfulness of local explanations produced by these methods.

These initial results validate the core components of our framework and demonstrate its potential to provide objective, nuanced, and comparable evaluations of local XAI methods, addressing a significant gap identified in the literature. Further experiments involving the OPS component and additional XAI methods are planned as outlined in Section 6.

6. Expected Next Research Steps and Final Contribution

The next steps focus on completing the evaluation of the benchmarking framework and expanding its application:

- **OPS Module Evaluation:** Rigorously evaluate the OPS metric using the validation protocol (Sec 4), assessing its sensitivity and ranking ability using simulated explanation sets and applying it to outputs from LIME, Anchor, and potentially SHAP.
- **Extended XAI Method Experiments:** Apply the full framework (Datasets, XFA, OPS, Overall Score) to evaluate additional state-of-the-art local XAI methods (e.g., SHAP) to provide a broader comparative analysis.
- **Publications:** Disseminate the findings related to the synthetic datasets, XFA, and OPS through publications in relevant peer-reviewed journals and conferences.
- **Thesis Completion:** Structure and write the PhD thesis, incorporating the benchmarking framework design, evaluations, experiments, use cases, and conclusions.

The **expected final contributions** of this research are significant:

1. **A Clear Distinction between Explanation Fidelity and Presentation:** By focusing on objective fidelity assessment separate from user perception, we provide a foundation for evaluating the core accuracy of XAI methods.
2. **Novel Evaluation benchmarking technique for Tabular Data:** The first benchmarking technique specifically for local, feature attribution XAI on tabular binary classification, addressing key limitations of prior work.
3. **Standardised Synthetic Ground Truth Datasets:** A publicly available benchmark suite enabling objective, reproducible evaluation and comparison of local XAI methods.
4. **Explainability Fidelity Assessor (XFA):** A novel, quantitative, multi-faceted metric (completeness, conciseness, sensitivity) for assessing explanation faithfulness against ground truth.
5. **Optimum Explainability Score (OPS):** A novel metric quantifying objective explanation characteristics (simplicity, coverage) to guide tuning towards optimal explainability.
6. **Advancement of XAI Evaluation Methodology:** Providing a systematic approach that incorporates objective metrics, enables comparisons, and addresses the challenges of evaluating local explanations, ultimately fostering the development of more robust, reliable, and comparable XAI methods.

While this work focuses on tabular binary classification, the underlying principles of using synthetic ground truth and objective metrics could potentially be adapted for other tasks (e.g., regression, multi-class) in future research.

Declaration on Generative AI

During the preparation of this work, the author used Gemeni in order to: Grammar and spelling check.

References

[1] D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2017) 2.

[2] S. Mohseni, N. Zarei, E. D. Ragan, Multifaceted interpretability of machine learning models: A survey, ACM Computing Surveys (CSUR) 55 (2023) 1–38.

[3] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[4] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[5] Z. C. Lipton, The mythos of model interpretability, Queue 16 (2018) 30.

[6] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[7] L. Rizzo, L. Longo, A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning, in: Proc. of the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, 2018, pp. 138–149.

[8] G. Vilone, L. Rizzo, L. Longo, A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence, in: Proc. of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Republic of Ireland, volume 2771, CEUR-WS.org, 2020, pp. 85–96.

[9] G. Vilone, L. Longo, Development of a human-centred psychometric test for the evaluation of explanations produced by xai methods, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 205–232.

[10] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller, Explainable ai: interpreting, explaining and visualizing deep learning, Springer (2019).

[11] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, European conference on computer vision (2014) 818–833.

[12] M. T. Ribeiro, S. Singh, C. 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, 2016, pp. 1135–1144.

[13] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, 2017, pp. 4765–4774.

[14] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-Precision Model-Agnostic Explanations, Proc. of 32nd Conference on Artificial Intelligence (AAAI) (2018) 1527–1535.

[15] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, Statistic Surveys 16 (2022) 1–85.

[16] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Machine Learning and Knowledge Extraction 3 (2021) 615–661.

[17] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.

[18] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, Advances in Neural Information Processing Systems 32 (2019).

[19] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE transactions on neural networks and learning systems 28 (2016) 2660–2673.

[20] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, Eraser: A benchmark to evaluate rationalized nlp models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7114–7129.

[21] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[22] W. Saeed, C. Omlin, Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems 263 (2023) 110273.