What's meant by explainable model: A Scoping Review

Mallika Mainali[1] , **Rosina O Weber**[1,2]
[1]Information Science, Drexel University
[2]Computer Science, Drexel University
{mm5579, rosina}@drexel.edu

Abstract

We often see the term *explainable* in the titles of papers that describe applications based on artificial intelligence (AI). However, the literature in explainable artificial intelligence (XAI) indicates that explanations in XAI are application- and domain-specific, hence requiring evaluation whenever they are employed to explain a model that makes decisions for a specific application problem. Additionally, the literature reveals that the performance of post-hoc methods, particularly feature attribution methods, varies substantially hinting that they do not represent a solution to AI explainability. Therefore, when using XAI methods, the quality and suitability of their information outputs should be evaluated within the specific application. For these reasons, we used a scoping review methodology to investigate papers that apply AI models and adopt methods to generate post-hoc explanations while referring to said models as *explainable*. This paper investigates whether the term *explainable model* is adopted by authors under the assumption that incorporating a post-hoc XAI method suffices to characterize a model as *explainable*. To inspect this problem, our review analyzes whether these papers conducted evaluations. We found that 81% of the application papers that refer to their approaches as an *explainable model* do not conduct any form of evaluation on the XAI method they used.

1 Introduction

Methods that attempt to produce information in lieu of explanatory value to make AI methods more transparent are very popular. As of April 2023, Google Scholar has reported about 140,000 papers that contain the term *XAI* in their title [GoogleScholar, 2023b]. Due to this popularity, XAI has been the topic of multiple workshops at AI conferences such as AAAI [Tulli *et al.*, 2022] and IJCAI [Madumal *et al.*, 2021; Miller *et al.*, 2021] in the last few years. The papers demonstrating XAI methods describe a variety of topics related to how AI methods can explain their decisions. XAI methods vary in the type of information content they produce as output. The categories of information outputs are feature attributions (*e.g.*, SHAP [Lundberg and Lee, 2017], LIME [Ribeiro *et al.*, 2016], DeepLIFT [Shrikumar *et al.*, 2019]), counterfactuals (*e.g.*, DiCE [Mothilal *et al.*, 2020], MACE [Yang *et al.*, 2022]), instance attributions (*e.g.*, influence functions [Koh and Liang, 2017], representer points [Yeh *et al.*, 2018], HYDRA [Chen *et al.*, 2021]), example-based (*e.g.*, [Nugent and Cunningham, 2005; Keane and Kenny, 2019]), and method-specific outputs such as decision tree paths (*e.g.*, [Izza *et al.*, 2022]).

Among these multiple XAI methods, SHAP [Lundberg and Lee, 2017] and LIME [Ribeiro *et al.*, 2016] have easy-to-use libraries that help them be quite popular. As of May 2023, the papers proposing SHAP and LIME have been cited, respectively, 12,740 [GoogleScholar, 2023a] and 12,504 [GoogleScholar, 2023d] times, while papers referencing instance attribution like influence functions have only been cited 2,193 times [GoogleScholar, 2023c]. Although these feature attribution methods are popular and easy to use, using these libraries under the assumption that the problem of explaining a model's decision is solved is not aligned with the body of literature that demonstrates their limitations and variability, and with those that prescribe domain- and application-specific evaluations [Lin *et al.*, 2020; van der Waa *et al.*, 2021; Rosenfeld, 2021a; Zhou *et al.*, 2021c; Rong *et al.*, 2022; Marques-Silva and Ignatiev, 2022]. Using these libraries with the claim of explainability without proper evaluation can be misleading because these feature attribution methods have been found to select misleading non-correlating features [Zhou *et al.*, 2021c].

Despite the popularity of many XAI methods, the literature often mentions criticisms of these methods (*e.g.*, [Adebayo *et al.*, 2018; Zhou *et al.*, 2021c; Marques-Silva and Ignatiev, 2022; Huang and Marques-Silva, 2023]). Said criticisms justify these methods be carefully evaluated to determine their suitability to specific application contexts, particularly because it is consensual that explanations are application-specific and contextual (*e.g.*, dependent on user, domain, and task [Hoffman *et al.*, 2017; Doshi-Velez and Kim, 2017; Gunning *et al.*, 2019; Mueller *et al.*, 2019; Barredo Arrieta *et al.*, 2020; Zhou *et al.*, 2021a]).

Another aspect to consider is that the output produced by feature attribution methods is the contributions of the features in each decision. This is considered important information with explanatory value but it is only one of many categories of information outputs that can be produced. They are not sufficient to describe the complete decision making process of a model. This limited scope combined with their variability in quality is yet another reason to include application-specific evaluations as many domains may be considered high-stakes such as healthcare, finance, privacy and security (*e.g.*, [Rudin, 2019; Marques-Silva and Ignatiev, 2022; Rudin *et al.*, 2022].

When examining publications using the term *explainable* to describe machine learning (ML) methods, and particularly referring to such methods as *explainable models*, we noticed that many of those papers implemented one of the popular XAI libraries without any evaluation. By not evaluating these XAI methods, these works are sending a misleading message that these XAI methods have solved the explainability problem and simply generating feature attributions makes their model *explainable*. Based on the literature as mentioned above, there needs to be a scientific evaluation of the applications that use XAI methods to confirm if those applications produce reliable explanations for the specific application and domain where the models are applied. This is because existing XAI methods can often be wrong, and can possibly even be dangerous if they are wrong when used in safety-critical applications [Rudin, 2019; Rudin *et al.*, 2022]. If these methods are being used to explain the decisions of AI models, the explanations need to be evaluated to ensure they are reliable in all applications, particularly high-stakes and safety-critical applications.

To get a scientific perspective on how often the XAI methods are evaluated when used to make a model explainable, in this paper, we employed a scoping literature review methodology. With the scoping review, we aim to gather scientific evidence on the use of the term *explainable model* with and without the presence of evaluation of XAI methods. To achieve this, this study will address the following research question (RQ):

RQ: How extensive is the practice of referring to a model as explainable simply because methods to provide feature attributions were used without evaluation of the explanations generated by the method?

In our study, we categorize an XAI method as evaluated if the authors of these papers conduct any qualitative or quantitative scientific experiment (*e.g.*, case studies, computational evaluation using cross-validation, fidelity test) to assess the quality of feature attributions produced by adopted AI methods. The next section of this paper describes the background of XAI methods and their evaluation approaches. Section 2.1 describes the advancements and limitations of feature attribution methods such as SHAP [Lundberg and Lee, 2017] and LIME [Ribeiro *et al.*, 2016]. Section 2.2 discusses the state of evaluation approaches for *post hoc* explanation methods. We present the the scoping review methodology in Section 3. In section 4, we show an analysis of the results and conclude with a discussion on the state of evaluation methods used for so-called *explainable models*.

2 Background

2.1 Background of XAI Methods

With the development of feature-attribution methods like LIME in 2016 [Ribeiro *et al.*, 2016] and SHAP in 2017 [Lundberg and Lee, 2017], model explainability has become an important aspect in designing black-box models. Recent trends show that the popularity of XAI methods has been growing exponentially over the last 5 years with the saturation in regards to deep learning [Angelov *et al.*, 2021]. While these methods do provide some interpretation of a model's predictions, the explanations generated may not always be informative, or accurate (*e.g.*, [Rudin, 2019; Rudin *et al.*, 2022; Marques-Silva and Ignatiev, 2022]). Multiple studies have pointed out that the use of SHAP or LIME libraries does not necessarily create an explanatory model and have dissected the potential limitations of these libraries [Hatwell *et al.*, 2021; Al-Merri and Ben Miled, 2022]. Since there is no formal definition of feature importance, it is difficult to validate these methods [Wilming *et al.*, 2021]. Because of these drawbacks, many researchers have been investigating the challenges and limitations of XAI methods, which stem from the explanation generation without the quantification of the explanation [Molnar *et al.*, 2020].

Angelov et al. dissected the use of different types of explanations, including SHAP and LIME, to demonstrate that the explanations generated by these models are not always efficient or reliable [Angelov *et al.*, 2021]. Similarly, Tan et al. have described the uncertainty of explanations [Tan *et al.*, 2019], and Rengasamy et al. have referenced the instability of LIME by citing its inability to give a global approximation of feature importance [Rengasamy *et al.*, 2021]. Additionally, multiple studies have attempted to analyze the explanations of the predictions made by two models to compare their soundness, consistency, and accuracy [Slack *et al.*, 2019; Alharbi *et al.*, 2021]. For instance, Alharbi et al. demonstrated the inconsistency between the explanations generated using SHAP of two models of the same image and proposed a novel approach to minimize the inconsistency [Alharbi *et al.*, 2021]. Similarly, a demonstration by Slack et al. revealed the weakness of *post hoc* explanation techniques by explaining how these methods could be easily fooled with adversarial classifiers [Slack *et al.*, 2019].

Some studies have showcased the need for the evaluation of XAI methods by demonstrating the inefficiency of the explanations generated by these methods after an evaluation with human experts [Rafferty *et al.*, 2022]. Rafferty et al. showed the unreliability of using *post hoc* explanation methods in high-stake decision-making by revealing that the explanations from the three techniques (*i.e.*, SHAP, LIME, and RISE) were unhelpful [Rafferty *et al.*, 2022]. This has led to the development of frameworks for explanations in sensitive fields like healthcare [Barda *et al.*, 2020], but a standardized framework for the evaluation of those explanations is still needed.

2.2 Evaluation of *post hoc* explanations

Despite the popularity of *post hoc* explanation methods, there are no standard methods to evaluate them. Some studies

& XAI methods evaluated & XAI methods not evaluated
New method of explanation proposed& 16& 33
Use of SHAP library& 10& 92
Use of LIME library& 3& 14
Combination of different XAI methods (SHAP and LIME)& 6& 13

Table 1: Classification of literature based on evaluation of XAI methods

have shown the necessity of developing a framework for the evaluation of XAI methods [Rawal *et al.*, 2022], and there have been works that proposed evaluation criteria for the quality of explanations (*e.g.*, [Sundararajan *et al.*, 2017; Alvarez-Melis and Jaakkola, 2018; Cui *et al.*, 2019; Montavon, 2019]). These criteria reiterate the need to evaluate an explanation based on different dimensions such as human comprehensibility, fidelity, accuracy, scalability, and generality [Dai *et al.*, 2022; Belle and Papantonis, 2021]. Despite the importance of an evaluation metric for XAI methods, there is currently no commonly recognized scale that is widely accepted by XAI researchers [Liu and Hu, 2022]. This could be because of the difficulty and the expense needed to quantify the explanations as Zhou et al. have demonstrated that the evaluation of the explanations is a multidisciplinary research area, which makes it difficult to define an implementation of a metric that can be applied to all applications of XAI methods [Zhou *et al.*, 2021b]. Nonetheless, Vu et al. have introduced x-Eval which is an evaluation metric that quantifies feature-based local explanation's quality [Vu *et al.*, 2019], and Cugny et al. have introduced AutoXAI, which provides a framework for automatically selecting an XAI method based on some evaluation metrics [Cugny *et al.*, 2022]. These attempts, however, have not been widely adopted.

3 Methodology

In this section, we describe the implementation o the scoping review methodology outlined by Arksey and O'Malley (2005). A scoping review collects and categorizes existing literature and identifies the nature and gaps of current research evidence. Contrary to systematic reviews that seek to answer pre-defined questions from a narrow range of quality-assessed evidence, scoping reviews examine the range of a research topic and identify the gaps within existing literature [Arksey and O'Malley, 2005]. The scoping review fits our goal, which is to examine the range of literature that use the term *explainable model* to indicate a model for which a *post hoc* XAI method was applied and to analyze the literature that demonstrates the lack of evaluation. Next, we describe data extraction and collection.

3.1 Data Extraction

The initial step in data extraction is to search the literature. Our primary focus was to identify papers describing the use of some model to execute an AI task such as classification in a given domain so that the papers can be characterized as *application papers* rather than *research papers*. It is typically when applying AI models within a specific application that users and consequently model explainability is considered. We had anecdotally observed that many papers were adopting some *post hoc* XAI method to generate feature attributions in lieu of explanation for those applications of models and referring to them as *explainable models*.

We executed the search targeting the period starting in January 2016 and going until December 2022. We started the search on the Scopus database for articles that contained the following keywords: explainable* OR explainability* OR interpretability* OR explainable approach* OR explainable method* OR explainable AI* OR explainable machine learning* OR XAI*. The other inclusion criteria were English language and application papers. To facilitate the search and the analysis, we included papers that referenced either of the two most commonly used feature attribution methods (*e.g.*, SHAP or LIME). This last criterion helped us select the right papers because of their popularity.

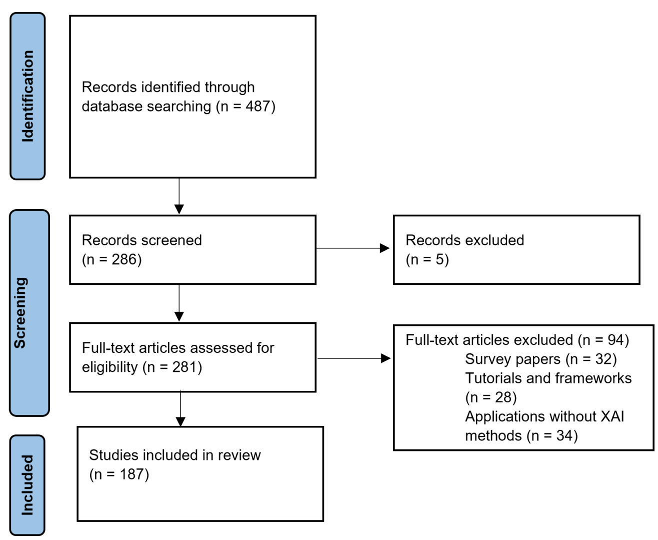

Figure 1: PRISMA flow diagram for the selection process of literature for the scoping review of the evaluation of XAI methods in application papers.

3.2 Data Collection

Figure 1 shows the steps Identification, Screening, and Inclusion describing how we moved from 487 initial papers to the final 187 papers. The screening process was conducted using the PRISMA extension [Page *et al.*, 2021] for scoping reviews (Figure 1). The titles and abstracts of the papers were screened to check for eligibility. The initial search in Scopus resulted in 487 articles that described some applications using XAI methods. These 487 articles were further filtered to 286 articles based on open-access availability. After the initial screening, all the data from the included articles were extracted for manual analysis. Out of 286 articles, five articles were excluded because they were not downloadable for data analysis. 281 full-text articles were downloaded and reviewed manually. The process of manual review consisted of this first author examining the full-text to ascertain the following aspects: 1. Whether the paper was indeed applying an AI model representing an application paper; 2. If the title or any other section described their applied model as *explainable* or *interpretable*; 3. If the paper really adopted a *post hoc* XAI method; and 4. If the XAI method was evaluated.

After this thorough review process, 94 articles were excluded. The reasons for excluding these 94 articles were: 32 articles were survey papers that did not describe any AI application, 28 articles were frameworks or tutorials, and 34 articles were papers that did not use any XAI methods, but only mentioned the term *explainable* in the paper without being used in an application. Finally, 187 articles qualified for final inclusion.

4 Results and Discussion

The resultant 187 articles were categorized either as evaluated or not evaluated. There were 35 application papers (19%) that evaluated the XAI methods and 152 application papers (81%) that did not evaluate them. Figure 2 shows a pie chart representing the percentage of articles that evaluated the adopted XAI method vs. articles that did not conduct any evaluation. 81% represents significant level of papers claiming to describe explainable models that do not evaluate the generated explanations. This number is even more concerning when we consider that a large number of those applications may be for high-stake decisions.

Within the 152 papers that did not evaluate the output of XAI methods, 33 articles included novel approaches. 14 articles utilized LIME libraries and 92 articles adopted SHAP. Finally, 13 articles included a combination of different libraries. Out of 187 application papers, 64% of the papers used SHAP and LIME feature attribution methods without any evaluation.

Within the category of evaluated XAI methods, the 35 papers found can be further categorized as follows. 16 articles proposed new evaluation methods. This might suggest that these articles were not merely evaluating the explanations of their application but they may have written the paper mostly for introducing the novel method. Only three articles that evaluated explanations had implemented LIME. 10 papers that evaluated their explanations had adopted SHAP. Finally, six articles used the evaluation step to compare the explanations between different XAI methods (*e.g.*, SHAP versus LIME).

Among the 16 articles that used a new evaluation method for their XAI approach, nine articles generated explanations using feature attribution methods, five utilized counterfactual explanations, one used a combination of counterfactual and example-based explanations, and one article utilized an XAI method to generate instance-based explanation. When we compared the evaluation approach of their proposed XAI methods, we found that five articles evaluated the stability

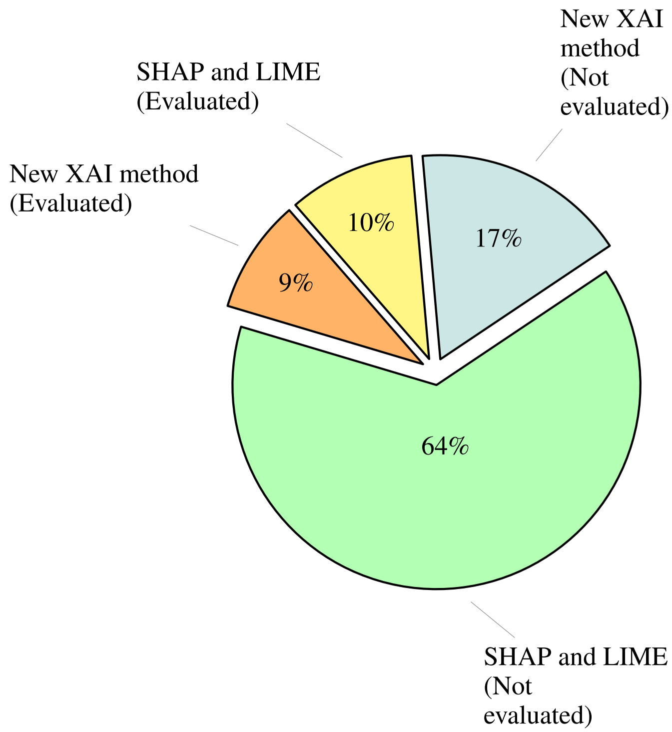

Figure 2: Pie chart showing a breakdown of evaluated and not evaluated XAI methods in reviewed papers

and faithfulness of the explanations produced by their model using some kind of metric. For example, Amich and Eshete introduced a novel evaluation metric for stability to evaluate the explanations of their model, EG-Booster. This stability metric enabled vetting the reliability of ML explanation methods before they were used to guide ML robustness evaluation [Amich and Eshete, 2021]. For this, they computed the similarity between the returned predictions of the same adversarial sample after EG-Booster attack using (k-l)-Stability and k-Stability metrics. Another novel approach to evaluate the explanation was introduced by Hatwell et al. for their algorithm of the black-box model AdaBoost. They measured the precision, coverage, and stability to evaluate the quality of the explanations generated by their model. To assess the stability and avoid over-fitting explanations, they used the formulation of (n+1)/(m+K), where n is the number of covered and correct instances, m is the number of covered instances, and K is the number of classes [Hatwell *et al.*, 2021]. Additionally, they assessed the coverage by evaluating the fraction of previously unseen instances a user could attempt to classify after seeing an explanation and precision by evaluating the fraction of those classifications that would be correct if a user applied the explanation correctly. Böhle et al. took a different approach to evaluate the faithfulness of their model by utilizing a grid-pointing game [Böhle *et al.*, 2022]. They evaluated the explanations generated by their model on a synthetic 3x3 grid of images of different classes and measured how much positive attribution an explanation method assigned to the correct location in the grid for each of the corresponding class logits.

Aside from evaluating the stability and faithfulness of the explanations, four articles used different versions of the fidelity metric to evaluate the explanations generated by their models. Panigutti et al. adopted the fidelity metric, Hit, and Explanation complexity to evaluate the different explanation pipelines of their model. As a way to calculate the fidelity to the black box $\in$ [0,1], they compared the predictions made by the interpretable model with the predictions made by the black-box model on a synthetic neighbourhood of the instance and measured the fidelity with micro-averaging [Panigutti *et al.*, 2020]. Similarly, Setzu et al. used fidelity $\in$ [0,1] to compare the predictions returned by the rules in a given explanation theory E or by the black box b [Setzu *et al.*, 2021]. Among the XAI methods that generated counterfactual explanations, three articles used different combinations of sparsity, validity, proximity, diversity, distance, and generation time to evaluate the quality of the explanations. For instance, Chen et al. referenced [Verma *et al.*, 2022] to define their evaluation criteria for validity, proximity, sparsity, and generation time of the counterfactual explanations [Chen *et al.*, 2022]. They defined validity as the ratio of the counterfactuals that meet the prediction goal to the total number of the counterfactuals generated, proximity as the distance of a counterfactual from the original input sample, sparsity as the number of features that must be changed according to a counterfactual, and generation time as the time required to generate counterfactuals. Crupi et al. used a similar metric to evaluate their proposed methodology to generate counterfactual explanations, called Counterfactual Explanations as Interventions in Latent Space (CEILS) by using a combination of validity, proximity, sparsity, and distance between the counterfactual and factual observation [Crupi *et al.*, 2021].

Although the XAI approaches utilized in the 16 articles were evaluated using some variation of quantitative evaluation, there was an evident lack of qualitative evaluation in these approaches. Since there is still ambiguity in the definition of explanation [Gilpin *et al.*, 2022; Buchholz, 2022], it is significant to place an emphasis on the rigorous qualitative evaluation of user studies [Johs *et al.*, 2020] and incorporate usability studies that can improve the user experience [Dieber and Kirrane, 2020]. Qualitative evaluation methods such as interviews, case studies, focus groups, and observations could be useful in understanding the success and failure of these approaches through a human-centered perspective. Furthermore, most papers that adopted certain evaluation methods did not reference any standard evaluation metrics. Referencing metrics such as [Hoffman *et al.*, 2019] that include the Explanation Goodness Checklist, Curiosity Check-list, Explanation Satisfaction Scale, and Trust Scale Recommended for XAI, or [Rosenfeld, 2021b] that is based on quantifying explainability based on the number of rules in the agent's explanation, the performance difference between the AI model and the explanation logic, the number of features used to create the explanation, and the stability of the explanation created by the model could be a useful starting point in facilitating a standard practice of evaluating XAI methods. The absence of human-centered and standard evaluation approaches suggest lack of rigour in the field.

5 Conclusions and Future Work

Through the scoping review of 187 application papers included in this study, we conclude that most models presented as explainable do not evaluate their XAI methods. This demonstrates a significant gap in publications of AI models that adopt XAI libraries. Furthermore, in the existing literature, evaluation of XAI methods is commonly included when authors propose a new method, or if the study is comparing two or more methods of explanation. We reiterate that XAI methods should be evaluated when they are adopted to produce explanations for a model. Researchers should also be careful when using XAI methods to avoid implying that existing libraries such as SHAP and LIME solve the problem of AI explainability.

In our literature review, we examined the frequency of the practice where papers use the term *explainable model* without evaluating the adopted explanation methods. The main limitation of our study is that it is not an exhaustive review of evaluations of XAI methods which becomes an opportunity for future work.

References

[Adebayo *et al.*, 2018] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[Al-Merri and Ben Miled, 2022] Mohammad Al-Merri and Zina Ben Miled. Global translation of classification models. *Information*, 13(5), 2022.

[Alharbi *et al.*, 2021] Raed Alharbi, Minh N. Vu, and My T. Thai. Learning interpretation with explainable knowledge distillation. *CoRR*, abs/2111.06945, 2021.

[Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. *CoRR*, abs/1806.07538, 2018.

[Amich and Eshete, 2021] Abderrahmen Amich and Birhanu Eshete. Eg-booster: Explanation-guided booster of ml evasion attacks, 2021.

[Angelov *et al.*, 2021] Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

[Arksey and O'Malley, 2005] Hilary Arksey and Lisa O'Malley. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8(1):19–32, 2005.

[Barda *et al.*, 2020] Amie J. Barda, Christopher M. Horvat, and Harry Hochheiser. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making*, 20(1):257, 2020.

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

[Belle and Papantonis, 2021] Vaishak Belle and Ioannis Papantonis. Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4, 2021.

[Buchholz, 2022] Oliver Buchholz. A means-end account of explainable artificial intelligence, 2022.

[Böhle *et al.*, 2022] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability, 2022.

[Chen *et al.*, 2021] Yuanyuan Chen, Boyang Li, Han Yu, Pengcheng Wu, and Chunyan Miao. Hydra: Hypergradient data relevance analysis for interpreting deep neural networks, 02 2021.

[Chen *et al.*, 2022] Ziheng Chen, Fabrizio Silvestri, Jia Wang, He Zhu, Hongshik Ahn, and Gabriele Tolomei. Relax: Reinforcement learning agent explainer for arbitrary predictive models, 2022.

[Crupi *et al.*, 2021] Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, and Beatriz San Miguel Gonzalez. Counterfactual explanations as interventions in latent space, 2021.

[Cugny *et al.*, 2022] 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*, CIKM '22, page 315–324, New York, NY, USA, 2022. Association for Computing Machinery.

[Cui *et al.*, 2019] Xiaocong Cui, Jung Min Lee, and J Hsieh. An integrative 3c evaluation framework for explainable artificial intelligence. In *Proceedings of the AI and Semantic Technologies for Intelligent Information Systems*, 2019.

[Dai *et al.*, 2022] Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, jul 2022.

[Dieber and Kirrane, 2020] Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime, 2020.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[Gilpin *et al.*, 2022] Leilani Gilpin, Andrew Paley, Mohammed Alam, Sarah Spurlock, and Kristian Hammond. "explanation" is not a technical term: The problem of ambiguity in xai, 06 2022.

[GoogleScholar, 2023a] GoogleScholar. Google scholar citations for "a unified approach to interpreting model predictions'. 2023.

[GoogleScholar, 2023b] GoogleScholar. Google scholar citations for titles with the term "xai", 2023.

[GoogleScholar, 2023c] GoogleScholar. Google scholar citations for "understanding black-box predictions via influence functions", 2023.

[GoogleScholar, 2023d] GoogleScholar. Google scholar citations for "why should i trust you?": Explaining the predictions of any classifier", 2023.

[Gunning *et al.*, 2019] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, 2019.

[Hatwell *et al.*, 2021] Julian Hatwell, Mohamed Medhat Gaber, and R. Muhammad Atif Azad. gbt-hips: Explaining the classifications of gradient boosted tree ensembles. *Applied Sciences*, 11(6), 2021.

[Hoffman *et al.*, 2017] Robert R. Hoffman, Shane T. Mueller, and Gary Klein. Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, 32(4):78–86, 2017.

[Hoffman *et al.*, 2019] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019.

[Huang and Marques-Silva, 2023] Xuanxiang Huang and Joao Marques-Silva. The inadequacy of shapley values for explainability. *arXiv preprint arXiv:2302.08160*, 2023.

[Izza *et al.*, 2022] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On tackling explanation redundancy in decision trees. *Journal of Artificial Intelligence Research*, 75:261–321, 2022.

[Johs *et al.*, 2020] Adam J Johs, Denise E Agosto, and Rosina O Weber. Qualitative investigation in explainable artificial intelligence: A bit more insight from social science. *arXiv preprint arXiv:2011.07130*, 2020.

[Keane and Kenny, 2019] Mark T. Keane and Eoin M. Kenny. How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In *Case-Based Reasoning Research and Development*, pages 155–171. Springer International Publishing, 2019.

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

[Lin *et al.*, 2020] Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors, 2020.

[Liu and Hu, 2022] Qian Liu and Pingzhao Hu. Extendable and explainable deep learning for pan-cancer radiogenomics research. *Current Opinion in Chemical Biology*, 66:102111, 2022.

[Lundberg and Lee, 2017] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 2017.

[Madumal *et al.*, 2021] Prashan Madumal, Silvia Tulli, Rosina Weber, and David Aha. Ijcai workshop on explainable artificial intelligence (xai), 2021.

[Marques-Silva and Ignatiev, 2022] Joao Marques-Silva and Alexey Ignatiev. Delivering trustworthy ai through formal xai. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12342–12350, Jun. 2022.

[Miller *et al.*, 2021] Tim Miller, Rosina Weber, and Ofra Amir. Ijcai 2022 workshop on explainable artificial intelligence (xai), 2021.

[Molnar *et al.*, 2020] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops*, pages 417–431. Springer International Publishing, 2020.

[Montavon, 2019] Grégoire Montavon. Gradient-based vs. propagation-based explanations: An axiomatic comparison. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 253–265. Springer International Publishing, Cham, 2019.

[Mothilal *et al.*, 2020] 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*. ACM, jan 2020.

[Mueller *et al.*, 2019] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai, 2019.

[Nugent and Cunningham, 2005] Conor Nugent and Padraig Cunningham. A case-based explanation system for black-box systems. *Artif. Intell. Rev.*, 24:163–178, 10 2005.

[Page *et al.*, 2021] 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, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

[Panigutti *et al.*, 2020] 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*, FAT* '20, page 629–639, New York, NY, USA, 2020. Association for Computing Machinery.

[Rafferty *et al.*, 2022] Amy Rafferty, Rudolf Nenutil, and Ajitha Rajan. Explainable artificial intelligence for breast tumour classification: Helpful or harmful. In Mauricio Reyes, Pedro Henriques Abreu, and Jaime Cardoso, editors, *Interpretability of Machine Intelligence in Medical Image Computing: 5th International Workshop, iMIMIC 2022, Held in Conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings*, Lecture Notes in Computer Science, pages 104–123. Springer, Cham, October 2022.

[Rawal *et al.*, 2022] 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, 2022.

[Rengasamy *et al.*, 2021] Divish Rengasamy, Benjamin C. Rothwell, and Grazziela P. Figueredo. Towards a more reliable interpretation of machine learning outputs for safety-critical systems using feature importance fusion. *Applied Sciences*, 11(24), 2021.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. 2016.

[Rong *et al.*, 2022] Yao Rong, Tobias Leemann, Thai trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: User studies for model explanations, 2022.

[Rosenfeld, 2021a] Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 45–50, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems.

[Rosenfeld, 2021b] Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 45–50, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems.

[Rudin *et al.*, 2022] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

[Rudin, 2019] 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.

[Setzu *et al.*, 2021] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx - from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457, 2021.

[Shrikumar *et al.*, 2019] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019.

[Slack *et al.*, 2019] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. How can we fool LIME and shap? adversarial attacks on post hoc explanation methods. *CoRR*, abs/1911.02508, 2019.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[Tan *et al.*, 2019] Hui Fen Tan, Kuangyan Song, Madeilene Udell, Yiming Sun, and Yujia Zhang. Why should you trust my interpretation? understanding uncertainty in LIME predictions. *CoRR*, abs/1904.12991, 2019.

[Tulli *et al.*, 2022] Silvia Tulli, Prashan Madumal, Mark T. Keane, and David W. Aha. Ijcai 2022 workshop on explainable artificial intelligence (xai), 2022.

[van der Waa *et al.*, 2021] 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.

[Verma *et al.*, 2022] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review, 2022.

[Vu *et al.*, 2019] Minh N. Vu, Truc D. T. Nguyen, NhatHai Phan, Ralucca Gera, and My T. Thai. Evaluating explainers via perturbation. *CoRR*, abs/1906.02032, 2019.

[Wilming *et al.*, 2021] Rick Wilming, Céline Budding, Klaus-Robert Müller, and Stefan Haufe. Scrutinizing xai using linear ground-truth data with suppressor variables, 2021.

[Yang *et al.*, 2022] Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven C. H. Hoi. Mace: An efficient model-agnostic framework for counterfactual explanation, 2022.

[Yeh *et al.*, 2018] Chih-Kuan Yeh, Joon Sik Kim, Ian E. H. Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks, 2018.

[Zhou *et al.*, 2021a] 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), 2021.

[Zhou *et al.*, 2021b] 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), 2021.

[Zhou *et al.*, 2021c] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *CoRR*, abs/2104.14403, 2021.