Investigation and Comparison of Evaluation Methods of Model-Agnostic Explainable AI Models

Vanisha Oedayrajsingh Varma
Supervisor(s): Chhagan Lal, Mauro Conti
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Many artificial intelligence (AI) systems are built using black-box machine learning (ML) algorithms. The lack of transparency and interpretability reduces their trustworthiness. In recent years, research into explainable AI (XAI) has increased. These systems are designed to tackle common ML issues such as trust, accountability, and transparency. However, research into the evaluation of XAI is still low. In this paper, common trends in the evaluation of state-of-the-art model-agnostic XAI models and any missing or undervalued evaluation methods are identified. First, a taxonomy is explored, and an overview of existing evaluation metrics found in literature is made. Then, using this overview, a thorough analysis and comparison of the evaluation methods of 5 state-of-the-art model-agnostic XAI models (LIME, SHAP, Anchors, PASTLE, and CASTLE) is done. It has been discovered that only a small subset of the found evaluation metrics is used in the evaluation of the state-of-the-art models. Metrics that are not often assessed in user-studies but deserve more attention are *(appropriate) trust*, *task time length*, and *task performance*. For synthetic experiments, only *fidelity* is commonly assessed. The models are also only assessed using proxy tasks, none of them are assessed using real-world tasks. In addition, each identified metric was found to have various different measurement methods and units of measurement, indicating a lack of standardization.

1 Introduction

For many, artificial intelligence (AI) has become a major part of everyday life. With the increasing prevalence of AI in social media, advertisement, product and entertainment recommendations, customer service, recruitment and hiring, self-driving cars, surgeries, and others, some might consider AI inescapable, indispensable even [1].

At the basis of these AI systems often lie black-box machine learning (ML) algorithms. These algorithms have hidden internal functionalities, which causes many decisions and predictions made by them to often not be well understood by their users. This lack of transparency and interpretability reduces their trustworthiness, which is especially critical in the use of ML techniques in high-stakes applications that greatly affect human lives, most prominently in the domains of medicine, criminal justice, transport and finance [2].

Explainable AI (XAI) attempts to improve on the uncertainty brought by ML algorithms by providing "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future" [3]. XAI techniques try to tackle common ML issues such as trust, accountability, and transparency by either the design of transparent models that are inherently interpretable, or by providing post-hoc explanations for a model's decisions and predictions. These explanations include natural language explanations, visual explanations, numeric explanations, and explanations by example [1][4].

However, trustworthiness does not only rely on the availability of explanations, but also on the quality of these explanations. Despite this, explanations produced by XAI are often not systematically assessed, nor is there a consensus on how to assess the validity and reliability of an explanation [5][6]. And even though in recent years there has been a vast increase in research into new XAI techniques, research into the evaluation of explanation generation techniques and the evaluation of generated explanations is still lacking behind [7]. This paper aims to fill some of the gaps in the research into the evaluation of XAI, with a focus on the set of model-agnostic XAI techniques. Specifically, the goal is to *identify common trends in the evaluation of state-of-the-art model-agnostic XAI models, and identify any missing or undervalued evaluation methods*. To this point, the following sub-questions are addressed:

- What different evaluation methods exist in the field of XAI?
- How are state-of-the-art model-agnostic XAI models evaluated?
- What evaluation methods are still missing or are undervalued in the evaluation of model-agnostic XAI?

The rest of this paper is organised as follows. First, a brief introduction to model-agnostic XAI is given in Section 2. In Section 3, a taxonomy and description of current evaluation methods of XAI can be found. Next, in Section 4, the evaluation of selected state-of-the-art model-agnostic XAI is analysed and compared. Section 5 then reflects on the ethics of this research. Section 6 provides a discussion on the results found in Section 4, and any missing evaluation methods or methods that deserve more attention are identified, along with recommendations for future work. Lastly, final conclusions are presented in Section 7.

2 Background

XAI can be in the form of inherently *interpretable models*, meaning no additional methods are needed to explain them. The models are either simple enough to be directly understood, can be split up into understandable parts, or are mathematically well understood [8]. Some examples are decision trees, linear and logistic models, and Bayesian methods [8]. Models that are not inherently interpretable can be explained using *post-hoc* explanations (e.g., natural language, visual, numeric, or example-based explanations [1][4]). Post-hoc techniques make use of (interpretable) helper models to generate explanations, which are called *surrogate* or *proxy* models. These surrogate models either fully mimic the behaviour of the model that is being explained, or approximate a sub-aspect of it.

Post-hoc XAI techniques can be categorised into *model-agnostic* and *model-specific* techniques. Model-agnostic techniques can be applied to any type of ML model, whereas model-specific techniques can only be applied to a particular type of ML model (e.g., neural networks) [9].

Ribeiro et al. argue that an explainer should be model-agnostic [10]. Often, users have to decide between a number of competing models for the same end-goal. Model-agnostic techniques allow users to generate the same type of explanations for each model, which allows for a more straightforward comparison of the models. In addition, model-agnostic techniques will have the ability to explain any future models.

Taking the importance of model-agnosticism into account, and the scope of model-agnostic and model-specific techniques and the potentially different ways they can be evaluated, this study focusses on the subset of model-agnostic techniques. For this research, several state-of-the-art model-agnostic models were selected, of which the evaluation methods will be analysed. These models are briefly summarized in Table 1.

Table 1: State-of-the-art model-agnostic XAI models

Model	Explanations based on	Year
LIME [10]	Local interpretable approximations of complex models, and identification of the local important variables.	2016
SHAP [11]	Assigning an importance value to each feature for a specific prediction.	2017
Anchors [12]	If-then rules called 'anchors', if an anchor holds, the prediction is (almost) always the same.	2018
PASTLE [13]	Feature importance combined with pivots.	2021
CASTLE [14]	Feature importance combined with clusters.	2021

3 Evaluation Methods of XAI

Before XAI techniques can be evaluated, there need to be precise definitions of the criteria that an XAI technique can be assessed on. Doshi-Velez and Kim [15] proposed the following division of evaluation approaches: application-grounded, human-grounded, and functionality-grounded. Each of these are detailed in the below paragraphs, and for each a list of relevant evaluation metrics found in literature is composed.

The lists are aimed to be as comprehensive as possible, but are by no means complete. Some metrics were vaguely described and had no measurement methods. This is mentioned where relevant, and these metrics will not be considered further. Metrics were also often found to have different names in various studies. Metrics with similar definitions are grouped, with alternative names given where applicable. An overview of all metrics can be found in Table 2.

3.1 Human-centred methods: application-grounded evaluation & human-grounded evaluation

Application-grounded and human-grounded evaluation approaches both involve user-studies, with the biggest difference being that the first involves end-users or domain-experts and the latter involves lay humans. Because of this similarity, [6] merges them into one group called *human-centred evaluation*. [16] also does not distinguish between application- and human-grounded evaluation, but instead considers only the *user aspect*. In this paper a similar approach is taken, as the metrics found are relevant for both evaluation types. All metrics can be measured for both end-users and lay humans, albeit in different contexts (end-task or proxy task).

Application-grounded evaluation approaches assess the quality of explanations within the context of their final application (i.e., the real-world objective the XAI system is designed for). These evaluation methods involve human experiments with end-users or domain experts, and are aimed to discover to what extent the explanations assist in the execution of the end-task. Because application-grounded approaches directly test the capabilities of the system for the real-world application that they are built for, they are a strong indicator of the quality of the explanations. A few drawbacks are that these evaluation methods are often (1) expensive, due to the need to compensate highly trained domain experts and (2) time consuming, due to the need for approval (e.g., by Human Research Ethics Committees) and the time needed to conduct the experiments [15].

Human-grounded evaluation measures the quality of explanations through simpler human experiments compared to application-grounded evaluation. These evaluation methods are carried out with lay humans, and with simplified tasks that still maintain the essence of the final application. These are aimed at testing more general concepts related to the quality of explanations. Though human-grounded evaluation is still time consuming for the same reasons as application-grounded evaluation is, it is less expensive due to the use of lay humans and still gives a strong indication of the quality of explanations [15].

The metrics to be used in human-centred evaluation are as follows:

Trust can be seen as the confidence a user has in the system. The subjective nature of trust means it is often measured through subjective methods, such as questionnaires and interviews [17]. Objective measures include user compliance and user perceived system competence. Trust can also be expressed as the degree to which a user agrees with the model, giving way to another objective measure: the percentage of predictions, both correct and incorrect, that a user accepts [16]. A user can then be considered to have 100% trust in a model when they accept all correct and incorrect predictions. However, total trust should not be desired unless the model is 100% accurate. To reduce the risk of users accepting wrong predictions, trust should not be higher than the model's accuracy.

Appropriate trust Regardless of how satisfying explanations are, users should be able to detect erroneous predictions. Thus, from the definition of trust follows appropriate trust [16], which can be defined as the user's ability to distinguish between correct and incorrect predictions. Appropriate trust is measured by having users identify correct and incorrect predictions. It is also of-

Table 2: Overview of evaluation metrics

Evaluation approach	Both model-agnostic & model-specific	Model-specific only
Human-centred	Trust, Appropriate Trust, Satisfaction, Understanding, Task time length, Task performance, Ability to detect errors/Bias detection, Physiological indicators, Preference	-
Functionality-grounded	Fidelity, Accuracy, Level of (dis)agreement, Reliability, Privacy, Agreement, Monotonicity, Non-sensitivity, Effective complexity, Consistency, Validity, Proximity, Sparsity, Diversity, Closeness, Feasibility, Identity, Separability, Novelty, Representativeness	Implementation invariance, Continuity, Selectivity

ten called the user's *accuracy* [16] or *human judgement* [8]. The other side of this coin is *persuasiveness*, which is the system's capability to nudge a user into a certain direction [8]. In certain contexts, such as recommender systems, false-positives and false-negatives are not undesirable, as their purpose is to convince users to try or buy something. Persuasiveness is thus often in conflict with appropriate trust.

Satisfaction tests the clarity and usefulness of an explanation based on the views of the end-users [4]. Satisfaction is measured using subjective methods, such as interviews, self-reports, and case studies, or quantified using Likert-scale questionnaires [17].

Understanding The ability for a human to grasp the workings of a model [6]. Mental models form a representation of the user's understanding, and are often measured using interviews, self-explanation, or Likert-scale questionnaires [17]. An objective method is to measure user's accuracy in predicting model output [4][17].

Task time length Also known as *response time* or *(time) efficiency* [8][18]. The time needed for a user to complete a proxy task. It gives an indication of how long it takes to build a viable mental model after being presented with an explanation [8]. An example could be, when given an explanation and an input, the time needed for a user to make a prediction of the model output [18]. Faster user decisions indicate an intuitive understanding of explanations [19], which is especially important in time-sensitive applications such as recommender systems and automated driving [8].

Task performance measures whether explanations improve user's decision making for a designated task [4]. Measurement methods thus depend on the specific task, which can either be a proxy task or an end-task. An example is the number of correct or incorrect diagnoses made by a doctor who is guided by a medical assistant system [8].

Ability to detect errors [19] and Bias detection The degree to which explanations help in detecting errors. This is closely related to *bias detection*: If users can detect erroneous predictions, they can identify general patterns of bias in the system by taking into account the frequency of the errors [16].

Physiological indicators Though there has not been a lot of research into this area yet, physiological indicators of humans could be measures of explanation quality. Zhou et al. [20] found that presenting users with the influence of training data points on predictions causes significant differences in Galvanic Skin Response (GSR) and Blood Volume Pulse (BVP), suggesting that these physiological signals can be used as indicators of user trust.

Preference and Confidence *Preference* was mentioned in [19]. Although there was no clearly formulated description, preference could either be the user's preference for a certain type of explanation, or their preference for a specific method. One could argue that preference is already implicitly measured through the aforementioned metrics. For example, higher trust in one method over the other could indicate a preference for that method. However, preference can also explicitly be determined through binary choice. *Confidence*, or more precisely the user's confidence in a system, was more often mentioned in [6], [17] and [19]. There were also no clear definitions for confidence, but it often seemed to be paired with trust, or even be used interchangeably with trust. For these reasons, from henceforth, confidence and trust will be considered the same metric.

Metrics that were found but had no clear definition and measurement methods were: *likelihood to deviate* [19], *expectation*, *curiosity*, *cognition*, *context knowledge* [16], and *reliance* [6][16][17].

3.2 Functionality-grounded evaluation

Unlike the aforementioned approaches, functionality-grounded evaluation approaches do not involve humans. Instead, these measure formal properties of the explainer as proxies for explanation quality, and are all quantitative metrics. Because these do not involve humans, they are not expensive nor time consuming, unlike the preceding approaches. In addition, these can be used earlier in the evaluation cycle when techniques are not yet complete, or when human experiments are deemed to be unethical [15]. What proxies to use often depends on the type of explanations produced by the system, though also general metrics exist.

Before listing the metrics belonging to this category, it is important to state that objective assessment of explanation quality using quantitative metrics should not replace human-centred evaluation whenever possible. The perceived quality of explanations is user- and context-dependent, and thus inherently subjective [19]. Instead, the quantitative metrics can be used to help make a selection of explanations to be used

in user-studies, which in turn will lead to more efficient user-studies. With that said, the functionality-grounded metrics can be found below.

Fidelity Also often called *faithfulness* or *soundness* [8][16]. Fidelity is a measurement for the correctness of explanations [17], and represents how accurately the explanation reflects the behaviour of the underlying model [16][21]. This metric is only relevant for post-hoc techniques, as the fidelity of inherently interpretable models is naturally always 100% [8]. Fidelity is considered to be one of the most important properties of an explanation, because a low fidelity would mean the explanation does not reflect the prediction at all, making the explanation essentially useless [21]. There are several methods to measure fidelity, there is not one defined standard. One method could be to determine the accuracy of the surrogate model output with respect to the output of the model that is being explained [18]. Another method is to compare generated explanations to golden standard explanations or explanations from some inherently interpretable model [17]. Researchers also consider consistency in explanation results and computational interpretability as evidence for explanation correctness. Some more measurements of fidelity will be discussed below.

Accuracy The fraction of predictions that are correct. This is a measure of correctness and only refers to the prediction quality of the surrogate model, and thus only applies to post-hoc techniques [8].

Level of (dis)agreement The fraction of instances that are (not) assigned the same label by an approximation and the black-box model it is approximating, and can be seen as a measure of fidelity [22].

Reliability and Privacy are vaguely described in [15] and [16], and no clear methods of measurement were found. *Privacy* means that sensitive information in the data is protected. *Reliability* seems to be closely related to *accuracy*, and could be considered synonymous with it [16].

Agreement Feature attribution (or feature importance) methods explain a prediction by calculating the contribution of each input feature toward the individual prediction. These contributions are the input features' importance values (attribution), which can be used to create a ranking of important features [23]. For these kind of explanations, *agreement* is a popular evaluation metric, which is a measure of the correlation between two ranked lists. Agreement is usually used to compare new explanation techniques to established ones. Two different methods are said to *agree* if there is a strong correlation between their computed importance rankings. [23] found two issues with this metric: First, even though two explanation methods can indicate the same features as important, if these features do not end up in similar positions in the rankings, the rankings will only be weakly correlated, and thus their agreement will still be low; Second, because explanations are task-, model-, and context-specific, there often is not a single ground-truth explanation. Because of this, agreement can only determine whether two explanations are similar, but not whether any of them are necessarily correct. For these reasons, [23] recommend practitioners to stop using agreement as an evaluation metric.

Monotonicity [24] A measurement for the faithfulness of a feature attribution explanation. It is defined as the Spearman's correlation between the absolute importance values and corresponding (estimated) expectations. The calculation for the expectations follows from the argument that the importance of a feature should be proportional to how imprecise the prediction would be if its value was unknown. A precise formula is given in [24].

Non-sensitivity [24] Another measurement indicative of the faithfulness of a feature attribution explanation. Ensures that an importance value of zero is only assigned to features to which the model is not functionally dependent on. A precise formula is given in [24].

Effective complexity [24] The minimum number of features (ordered on importance value) that are needed to meet a performance measure of interest. A low effective complexity means that some features can be ignored because their effect is small (non-sensitivity). Explanations with low effective complexity are said to be both simple and broad.

Implementation invariance, Continuity, Selectivity These metrics are defined for deep neural networks, and are thus model-specific metrics. *Implementation invariance* means that attributions for functionally equivalent (the outputs are equal for all inputs, despite different implementations) networks should always be identical [25]. To ensure that two nearly equivalent data points have explanations that are also nearly equivalent, *continuity* is defined as a quality metric [19]. *Selectivity* measures how fast the prediction value goes down when removing features with the highest attributions [19].

Consistency A generalisation of *implementation invariance* that is relevant for post-hoc techniques. States that functionally equivalent models should produce the same explanation [8].

For counterfactual explanations, [26] identified the most commonly used evaluation metrics. Counterfactual explanations tell the user what features need to be changed in order to transition to the desired outcome by using (close) datapoints that already have the desired outcome. An essential property is thus that these changes need to be *actionable* (i.e., it is actually possible to make changes to the suggested features). For example, changes suggested to immutable features such as race and country of birth are not actionable, whereas changes to mutable features such as income and education are actionable. In summary, the following quantifiable proxies for ease of actionability are defined:

Validity The ratio of counterfactuals with the desired class label to the total number of generated counterfactuals. Higher validity is preferred.

Proximity The distance from the input datapoint to the generated counterfactual. Counterfactuals that are closer to

the input datapoint are easier to act upon, because they require less and smaller changes to be made.

Sparsity The amount of feature change needed to get to the desired outcome. A small number of features is preferred, though there is no consensus on a hard cap.

Diversity When multiple counterfactuals are generated for a single datapoint, it is important that at least one is easily actionable for the user. To allow the user to choose the easiest one, a diverse set of counterfactuals is desired. Diversity is defined as the distance between each pair of counterfactuals, and higher diversity is preferred.

Closeness to training data The average distance to the k-nearest datapoints. Counterfactuals closer to the training data are seen as more realistic.

Feasibility Measures if modifications to a counterfactual are realistic by determining if they satisfy the causal relation between features. To be considered actionable and realistic, it is desired that counterfactuals maintain any known causal relations between features. For example, getting a new degree also means an increase in age, age cannot decrease.

In [21], several more functionality-grounded metrics for the correctness of explanations are proposed, though no clear measurement methods were given:

Identity states that identical input instances should have identical explanations. An instance that is explained multiple times by the same XAI method is expected to always get the same generated explanation. If this is not the case, the method can be considered inaccurate due to its random nature.

Separability states that non-identical input instances should not have identical explanations. Follows from the same logic as *identity*, but there is one caveat: Separability does not hold if the model has more degrees of freedom than needed to represent the prediction function.

Novelty states that input instances should not come from a region in instance space that is far from the distribution of the training data. If it is, the model can be inaccurate, making the explanation useless.

Representativeness describes how many instances are covered by an explanation. Explanations can cover the entire model or only an individual prediction.

4 Comparison of Evaluation Methods of State-of-the-Art Model-Agnostic XAI

Based on the previously identified evaluation methods listed in Section 3, an analysis of the evaluation of 5 state-of-the-art model-agnostic XAI models (see Table 1) will be done. A comparison of their evaluation approaches and metrics can be found in Table 3.

LIME

Ribeiro et al. [10] evaluated LIME with both user-studies and simulated experiments. In total, three simulated experiments and three user-studies were done. For the simulated experiments, none of them involved humans, and all can thus be considered functionality-grounded evaluation. In the first experiment, inherently interpretable models (sparse logistic regression and decision trees) were trained in such a way that they use at most 10 features for any instance. These features are the golden set of features that these models consider important. Then, the test set is used to generate predictions and corresponding explanations. The fraction of golden features that are present in each explanation is calculated and averaged over all test instances. The authors refer to this as the *recall on truly important features* and its goal is to measure the *faithfulness (fidelity)* of the explanations.

In the second simulated experiment, 25% of features were randomly selected to be 'untrustworthy'. First, using only a black-box classifier, test set predictions were labelled untrustworthy if removing the untrustworthy features from the input changed the prediction, and otherwise they were considered trustworthy. Then, using LIME, test set predictions were deemed untrustworthy if removing *all untrustworthy features that are present in the corresponding explanation* changes the prediction. The predictions that are deemed either trustworthy or untrustworthy by the black box and by LIME explanations are then compared, and the F1-score on the trustworthy predictions is measured. With this experiment, the authors assert that they are trying to determine if LIME is helpful in assessing trust in predictions. One could argue that they are doing this by measuring the *fidelity* of LIME explanations to the black box. In particular, the *level of (dis)agreement* seems to be measured.

In the last simulated experiment, noisy features are intentionally added to a dataset on which pairs of competing classifiers are trained. The classifiers are trained in such a way that between pairs the difference in validation accuracy is within 0.1% and the difference in test accuracy is at least 5%. The better classifier is the one with higher test accuracy, but it is not possible to determine this just by examining the validation accuracy. Explanations are generated for X instances from the validation set, and explanations that contain the noisy features are marked as untrustworthy. For each classifier, the total predictions in the validation set marked as untrustworthy are determined, and the one with the least untrustworthy predictions is the one selected as the better classifier based on explanations. This is compared to the actual better classifier based on test set accuracy. Multiple runs are done, and the accuracy of picking the correct classifier based on explanations is calculated over varying values of X. The goal of this experiment is to determine whether the explanations provided by LIME can assist in model selection. Though this can be considered functionality-grounded evaluation due to the lack of user-involvement, none of the prior identified metrics relate to the goal of this experiment.

In the first user-study, a classifier was trained on a flawed dataset; the dataset contained features that do not generalize. Another version of the same classifier was trained on a cleaned version of the dataset, where many features that do not generalize had been removed. The subjects of the study were lay humans recruited through Amazon Mechanical Turk. The users were then shown raw data side-by-side with explanations, and their accuracy in choosing the better

Table 3: Comparison of the evaluation of 5 model-agnostic XAI models

Model	Application-grounded evaluation	Human-grounded evaluation	Functionality-grounded evaluation	Real-world task	Proxy task	Human-centred metrics	Functionality-grounded metrics
LIME	✓	✓	✓		✓	Trust Understanding Task performance Task time length Ability to detect errors/Bias detection	Fidelity Level of (dis)agreement 1 unseen
SHAP		✓	✓		✓	1 unseen	Accuracy Fidelity Computational efficiency (unseen)
Anchors	✓		✓		✓	Understanding Task time length	Fidelity
PASTLE	✓				✓	Understanding 1 unseen	
CASTLE	✓		✓		✓	Understanding	Fidelity Computational efficiency (unseen)

algorithm (in this case the 'cleaned' classifier) was measured. Specifically, the users were asked to select what they thought would be the algorithm that would perform best in the real world. Explanations were either generated by LIME, or by a greedy procedure that served as a baseline. From the fact that the human subjects chosen were lay humans, this can be considered human-grounded evaluation. In this experiment, the authors tested whether the explanations provided by LIME established sufficient *trust* and *understanding* for the users to pick the better algorithm.

In the second user-study, lay humans were again the subjects. Here, the 'bad' classifier from the previous study (trained on uncleaned data) was used to show explanations in addition to raw data. The users were asked to mark features for removal based on the explanations. These were then removed from the dataset, which was used to train new classifiers. The experiment started with 10 subjects, after which multiple rounds were done. Each round, multiple new users marked features for removal and multiple classifiers were trained. For each round, a path of classifiers originating from the first 10 subjects could be traced, and for each path the average accuracy of the classifiers was measured on a real-world dataset. In addition, the average accuracy across all paths was also measured. For each subject, the time spent cleaning was also measured. This was again a human-grounded evaluation. Now, *task performance* and *task time length* were measured. The task here was to improve the model by performing feature engineering based on explanations, and (the improvement in) model accuracy was used as a measure to gauge its success.

The last user-study was done with graduate students that have taken at least one graduate ML course. A classifier was trained to distinguish between pictures of wolves and huskies. The classifier was intentionally made biased by hand selecting the training data. All pictures of wolves had snow in the background, whereas the husky pictures did not. This way, the classifier learned to predict 'wolf' when there is snow, and 'husky' otherwise. The users where then shown 10 predictions without explanations, of which 8 were classified correctly, 1 contained a wolf without snow and was thus predicted as 'husky', and 1 contained a husky with snow and was thus predicted as 'wolf'. The subjects where then asked whether they trusted the algorithm to work well on real world data, why they did or did not, and how they think the classifier distinguishes between wolves and huskies. Then, the users where shown the images and corresponding explanations, and were asked the same questions again. The users in this study were by no means lay humans, and can be considered domain-experts, making this application-grounded evaluation. Through interview questions, the effect of explanations on *trust* and *the ability to detect errors/bias detection* (could users identify snow as a potential feature) was measured.

SHAP

Lundberg and Lee [11] evaluate SHAP with three experiments. First, the number of function evaluations needed to get accurate feature importance estimates was compared between SHAP, Shapley sampling values, and LIME. Here, the goal was to evaluate the *computational efficiency* and *accuracy* of SHAP. Since this does not involve humans, it can be considered functionality-grounded evaluation. However, none of the prior metrics correspond to *computational efficiency*.

Then, in a user-study, SHAP was compared to two other feature attribution methods: LIME and DeepLIFT. The subjects in the study were lay humans recruited through Amazon Mechanical Turk. Explanations generated by each method were compared to explanations of simple models provided by the users. In particular, the *agreement* between user explanations and the generated explanations was measured. The goal was to evaluate whether the generated explanations where *consistent with human intuition of simple models*. Since this was a user-study with lay humans, it can be considered human-grounded evaluation. However, none of the prior identified human-grounded metrics can be related to this method.

Lastly, the performance of SHAP, DeepLIFT, and LIME on MNIST digit image classification was compared. The accuracy in predicting the correct classification based on the explanations was measured and compared. Since no humans were involved, this can be considered functionality-grounded evaluation. As predictions from explanations were compared to the expected output, this can be seen as a measure of *fidelity*.

Anchors

Ribeiro et al. [12] first evaluated Anchors through a simulated experiment. Three different models were trained on the same datasets, and explanations were generated using LIME and Anchors. Users were simulated, and coverage was measured as the fraction of instances they predict after seeing explanations, and precision was measured as the fraction of the predictions that were correct. Since no real users were actually involved, this can be considered functionality-grounded evaluation. Since predictions are made based on explanations and then verified against the expected output, this can be seen as a measure of *fidelity*. A high coverage and precision indicate that the explanations accurately reflect the output of the underlying model, and thus indicate a high fidelity. Though Anchors did better on both counts, the experiment was not very indicative of real users. Therefore, a user-study was done.

The user-study was done with students who had or were taking a ML course. The users were shown 10 predictions. First without explanations, then with one, and two LIME or anchor explanations. Before and after seeing each round of explanations, they had to predict the output of the classifier on 10 random unseen instances. Users were only supposed to predict if they were very confident, and answer "I don't know" otherwise. The coverage was measured as the fraction of instances where the users made an actual prediction (and not "I don't know"), and precision was only measured in those instances. In addition, the time spent per prediction was also measured. Lastly, a poll was used to ask user's preferences for explanations (Anchors vs LIME). Since the users had knowledge of ML, they are not considered lay humans. Thus, this can be classified as application-grounded evaluation. By measuring coverage and precision, user's *understanding* of the model was quantified. A combination of high coverage and precision indicates high understanding. *Task time length* was also measured to quantify how long it takes to understand the explanations. *Preference* was measured subjectively through the poll.

PASTLE

La Gatta et al. [13] evaluated PASTLE through two user-studies. For both user-studies, the subjects had basic knowledge about ML. Thus, both user-studies can be considered application-grounded evaluation. In the first experiment, the users were shown explanations for 5 instances, either generated by PASTLE or by LIME. Then, they had to sort the instances in ascending order to what they perceived to be the probability of the instance being assigned a predetermined class. The following metrics were then computed using the results:

- *Mean Precision (P@k)*: The fraction of correctly sorted instances among the first k of the ranking.
- *Mean Average Precision (mAP)*: Penalizes errors in the first positions of the ranking averaging the overall Average Precision (AP) value for each answer.

These metrics were used to quantify the user's *understanding* of the model after having seen explanations.

In the second experiment, a ground-truth ordering was obtained based on the probability of the instances being assigned the predetermined class by the black-box model. Similar to the previous experiment, users were asked to rank the instances after having seen PASTLE or LIME explanations. These rankings were compared to the ground-truth ranking. The aim of this experiment was to determine whether the differences between rankings was due to the different explanation techniques, or due to random variance. However, no prior found metric can be related to this aim.

CASTLE

CASTLE provides decision rules which suggest how the prediction of the model generalizes to unseen instances, and outputs local information about feature importance [14]. La Gatta et al. [14] evaluated CASTLE by comparing it to Anchors, which also provides decision rules (see Table 1), in a number of experiments. First, a computational comparison between CASTLE and Anchors was done by comparing the run times of both. Since no users are needed for this, it can be considered functionality-grounded evaluation. However, it cannot be related to any of the prior found metrics.

Then, explanations for the same test set where generated using both CASTLE and Anchors. For each decision rule, the coverage was measured as the fraction of instances that are covered by the rule. The precision was measured as the fraction of instances that are correctly classified by the rule. This experiment also did not involve humans, and is thus considered functionality-grounded evaluation. Similar to the previously analysed Anchors experiment, classifications are made based on the explanations and verified against the expected output. Thus, this can again be seen as a measure of *fidelity*, with high coverage and precision indicating high fidelity.

Lastly, a user-study was done with undergraduate ML and statistics students. The design is similar to the user-study performed in [12]. First, users were shown 10 predictions without explanations, after which they had to predict the class of 10 unseen instances. Then, users were shown 10 predictions with explanations, either generated by CASTLE or by Anchors, after which they again had to make predictions of 10

unseen instances. Users were asked to predict only if they were confident, and had to answer "I don't know" otherwise. Coverage and precision were calculated similarly to the user-study performed in [12] (refer to the analysis of Anchors). Since the subjects had knowledge of ML or statistics, this can be considered application-grounded evaluation. The coverage and precision measure were used to quantify the user's *understanding* of the model.

5 Responsible Research

Ethics are an integral part of any research. This paper has been written with ethical responsibility in mind. Most of this research consisted of studying existing literature, for which the references are listed below. All information taken from these sources is properly cited throughout the text, and no source has been misrepresented or misquoted, ensuring that this research is reproducible. To ensure reliability, nearly all sources are peer-reviewed scientific papers. Some sources were only available on *arXiv*[1], and where thus not peer-reviewed. However, before using them for the purposes of this paper, they have been checked to ensure they come from trusted researchers and institutions, and contain reliable information.

6 Discussion and Future Work

From the results in Section 4 and in particular Table 3, it can be seen that only a small subset of the identified metrics in Section 3 is actually used in the evaluation of the state-of-the-art model-agnostic XAI models. From the prior identified human-centred metrics, 6 out of 9 were found to be measured. In comparison, this was only 3 out of 20 for the prior identified functionality-grounded metrics. Thus, a wider variety of human-centred metrics is used in evaluation compared to functionality-grounded metrics.

For human-centred evaluation, the most measured metric is *understanding*, which is measured in the evaluation of 4 models. There is an apparent lack of *trust* assessment, despite this being one of the main goals of XAI. *Task time length* and *task performance* are practical metrics to quantify the effectiveness of explanations and to compare different XAI models, but seem to currently be overlooked as they are only measured for 2 models.

For functionality-grounded evaluation, *fidelity* appears to be the most measured metric, which is also measured for 4 models. As *level of (dis)agreement* is just another measure of *fidelity*, the only other prior identified functionality-grounded metric measured is *accuracy*. However, *accuracy* is only measured for one model, despite it being an effective measure of correctness. In Section 3, various metrics are listed to assess the quality of feature attribution methods. Despite 4 out of 5 models using some form of feature attribution (LIME, SHAP, PASTLE, and CASTLE), none of these metrics are assessed.

In addition, there are also metrics used in the evaluation of the 5 analysed XAI models that were not already identified in Section 3, which are indicated in Table 3 as *unseen*. There are four total, the first being *computational efficiency*, which is measured for both SHAP and CASTLE. Since it was not present in any of the surveys that were used to create the list in Section 3, it can be concluded that computational efficiency is not often measured in the evaluation of XAI. Higher computational efficiency can be desired in time-sensitive applications, such as recommender systems and automated driving, but also in less time-sensitive applications, as waiting for explanations to be generated can negatively impact the user experience. For SHAP *the consistency of generated explanations with human intuition* was also measured. The remaining two unseen metrics were found in the evaluation of LIME and PASTLE. For LIME, the accuracy of picking the correct classifier based on explanations was measured in a functionality-grounded experiment. In the evaluation of PASTLE, PASTLE and LIME were compared, and it was assessed whether different user rankings of features was due to the different explanation techniques or due to random variance.

Furthermore, from both the listed metrics in Section 3 and the analysis of the evaluation methods in Section 4, it is obvious that for each identified metric, there is a wide variety of measurement methods and units of measurement. If the identified metrics are to be used to compare different XAI models, it would be beneficial to have standardised measurement methods and units of measurement for each metric. This will allow for more straightforward comparisons. Some identified metrics in Section 3 also had no clear definitions. If these are to be used for evaluation, they should be clearly defined and have clear methods of measurement.

Lastly, it can also be noted from Table 3 that even though both human-centred and functionality-grounded evaluation is nearly always done (with the exception being for PASTLE), this is always done with proxy-tasks, and never in the context of real-world applications. To have a stronger indication of success of explanations, it would be beneficial to have more tests in real-world applications.

From the above observations, the following future research directions are proposed:

- Evaluate the state-of-the-art models on unmeasured or rarely measured human-centred metrics found in Section 3, most notably *(appropriate) trust*, *task time length*, and *task performance*.
- Evaluate the state-of-the-art models on unmeasured or rarely measured functionality-grounded metrics found in Section 3, such as *accuracy*, or the measures of correctness (*identity*, *separability*, *novelty*, *representativeness*).
- Evaluate LIME, SHAP, PASTLE, and CASTLE using the functionality-grounded metrics specific for feature attribution methods found in Section 3, such as *monotonicity*, *non-sensitivity*, and *effective complexity*.
- Assess the benefits and potential future use of the unseen methods and metrics found in the analysis in Section 4, such as *computational efficiency* and *the consistency of explanations with human intuition*.
- Clearly define and standardise measurement methods for all metrics in Section 3.

[1] https://arxiv.org/

- Evaluate the state-of-the-art model-agnostic XAI models using real-world tasks.

7 Conclusion

In this paper, an analysis of different evaluation methods of model-agnostic XAI was conducted in order to identify common trends and any missing or undervalued evaluation methods. A taxonomy was presented, and currently used evaluation metrics were divided into human-centred and functionality-grounded evaluation. The results show that there is a lack of variety in the metrics that are assessed in the evaluation of the state-of-the-art model-agnostic XAI models. *Understanding* and *fidelity* are the most commonly tested metrics, and there is an apparent lack of *trust* assessment in human-centred evaluation, despite this being one of the main goals of XAI. Other overlooked metrics in human-centred evaluation are *task time length* and *task performance*. For functionality-grounded evaluation, there is an even greater amount of unused but promising evaluation metrics. Specifically, several metrics were found to assess feature attribution methods. Despite LIME, SHAP, PASTLE, and CASTLE all being feature attribution methods, none of these metrics were assessed. In addition, several promising metrics for correctness were found: *identity*, *separability*, *novelty*, and *representativeness*. Lastly, there is a lack of evaluation using real-world tasks, and a lack of standardisation in measurement methods and units of measurement.

References

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[3] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[4] A. Rawal, J. Mccoy, D. B. Rawat, B. Sadler, and R. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, 2021.

[5] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018.

[6] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[7] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Applied Sciences*, vol. 12, no. 3, p. 1353, 2022.

[8] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts," *arXiv*, 2021.

[9] S. Das, N. Agarwal, D. Venugopal, F. T. Sheldon, and S. Shiva, "Taxonomy and survey of interpretable machine learning method," *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *32nd AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018.

[13] V. La Gatta, V. Moscato, M. Postiglione, and G. Sperlì, "Pastle: Pivot-aided space transformation for local explanations," *Pattern Recognition Letters*, vol. 149, pp. 67–74, 2021.

[14] ——, "Castle: Cluster-aided space transformation for local explanations," *Expert Systems with Applications*, vol. 179, p. 115045, 2021.

[15] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv*, 2017.

[16] H. Löfström, K. Hammar, and U. Johansson, "A meta survey of quality evaluation criteria in explanation methods," *Lecture Notes in Business Information Processing*, pp. 55–63, 2022.

[17] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 11, no. 3-4, pp. 1–45, 2021.

[18] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez, "An evaluation of the human-interpretability of explanation," *arXiv*, 2019.

[19] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021.

[20] J. Zhou, H. Hu, Z. Li, K. Yu, and F. Chen, "Physiological indicators for user trust in machine learning with influence enhanced fact-checking," *Lecture Notes in Computer Science*, pp. 94–113, 2019.

[21] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[22] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable explorable approximations of black box models," *arXiv*, 2017.

[23] M. Neely, S. F. Schouten, M. Bleeker, and A. Lucic, "A song of (dis)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing," *arXiv*, 2022.

[24] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," *arXiv*, 2020.

[25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 3319–3328.

[26] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv*, 2020.