

Study on the Helpfulness of Explainable Artificial Intelligence (XAI)

Tobias Labarta^{1,2}, Elizaveta Kulicheva², Ronja Froelian², Christian Geißler², Xenia Melman², and Julian von Klitzing^{1,2}

¹ Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany^{**}

² Technische Universität Berlin, 10587 Berlin, Germany

Abstract. Explainable Artificial Intelligence (XAI) is essential for building advanced machine learning-powered applications, especially in critical domains such as medical diagnostics or autonomous driving. Legal, business, and ethical requirements motivate using effective XAI, but the increasing number of different methods makes it challenging to pick the right ones. Further, as explanations are highly context-dependent, measuring the effectiveness of XAI methods without users can only reveal a limited amount of information, excluding human factors such as the ability to understand it. We propose to evaluate XAI methods via the user's ability to successfully perform a proxy task, designed such that a good performance is an indicator for the explanation to provide helpful information. In other words, we address the helpfulness of XAI for human decision-making. Further, a user study on state-of-the-art methods was conducted, showing differences in their ability to generate trust and skepticism and the ability to judge the rightfulness of an AI decision correctly. Based on the results, we highly recommend using and extending this approach for more objective-based human-centered user studies to measure XAI performance in an end-to-end fashion.

Keywords: Explainable Artificial Intelligence (XAI) · Goodness measures and evaluation · Human-Centered XAI · XAI User-Study · XAI Evaluation.

1 Introduction

Society is facing the exponential application and exploitation of deep neural network-powered artificial intelligence (AI) in various areas of daily life. AI systems that make use of learned models to solve classification, segmentation, and transformation tasks for different input modalities such as images, video, text, and natural spoken language have shown or have the potential to outperform humans on specific tasks [14,17]. Since they are used in domains such as mobility, energy management, finance, medical diagnosis, and in general health, security,

^{**} We express our gratitude to Fraunhofer Heinrich-Hertz-Institute for financially supporting our work.

and many further critical domains, risk management and reduction plays a crucial role in building safe AI systems, which is also increasingly required by recent regulatory advances [9, 12].

To analyze and validate AI systems despite their statistical and often non-deterministic process of creation and their complex structure [20], instead of analytical proofs, proxy methods such as empirical evaluations and explainable AI (XAI) are used. Empirical evaluations using test datasets and calculating performance measures such as a confusion matrix, f-scores or AUC provide a narrow view of what the model learned. A vast amount of very different explainable AI methods have been proposed to understand, question, and investigate the models' inner workings and decisions, not just during testing but also while they are being applied in a real context. Choosing the right set of them for a specific application context is hard, especially if that context requires human operators to stay in control of the AI system. Thus, it is essential to develop methods for measuring and comparing the performance of XAI methods, especially when applied within specific contexts aimed at achieving distinct human-centered objectives. We term this "human task-related performance helpfulness." In this framework, "helpfulness" is defined as a quantifiable improvement in user performance on tasks that are aligned with the goals facilitated by the provision of explanations. In our experiment, the primary function of the XAI is to aid individuals in making informed decisions about their trust in the AI's results.

Measuring effects on users is usually tested in a user study. Most of these studies focus on usability aspects such as satisfaction which are biased by user preferences and opinion. They are not general proof if a certain explanation has a positive effect on the safety-related performances of an AI system. They therefore measure trust (belief) related performance, not the ability to which the explanation helps to control the AI system.

In this paper, we address the following research questions:

1. How accurate can XAI methods enable a user to *judge* AI decisions?
2. How far can XAI methods enable a user to *trust* AI decisions?
3. How far can XAI methods enable a user to *question* AI decisions?

The rest of the paper is structured as follows: We start with an overview of the state of the art of measuring the performance of explainable AI methods. We continue by contributing a new approach for measuring XAI performance in human-centered user studies on objective performance criteria. We further demonstrate the approach by conducting a user study to compare six well-known XAI methods concerning their ability to enable the user to correctly judge the trustworthiness of an AI-based classification decision. Finally, we discuss the insights drawn by this human-centered, objective evaluation approach.

2 Measuring Explainability

2.1 Approaches for measuring explainability

Previous theoretical work on explainability and its measurement are often inspired by research from domains such as human-computer interaction, psychol-

ogy, philosophy, or machine learning. A common finding in most works is, that explanations are context dependent. Therefore, they are required to define the recipients of the explanation (the explainees), the use case in which they operate the AI system, and the specific situation in which the explanation is provided [6, 27, 43]. The typology of an explanation is given by an explanandum, the object (thing or phenomenon) to be explained, the explanans (the actual explanation that is perceived by the users), and the relationship between both [5]. Cabitza further provides the following broad definition for the explanation provided by an XAI-system: It is "*the output of any computational system aimed at making AI-generated advice more understandable, appropriable and exploitable by their intended users and decision makers*".

In contrast to this specific perspective, some suggest measuring the degree of completeness of an explanation by providing a set of all possible questions [26, 36, 39]. "A set of all possible questions" refers to a complete collection of questions that could be asked to evaluate and understand an AI's behavior and decisions fully. Effectively, this would cover all aspects of the AI's operation, rationale, outcomes, and underlying mechanisms to ensure that the explanation of the AI is complete. The idea behind providing such a set is to measure how well an explanation satisfies the informational needs of users concerning the AI system. If an explanation can address all possible questions, it can be assumed complete in the sense that it leaves no aspect of the AI unexplained. However, providing such a complete set of questions for a real, practical use case is challenging as it requires interpretation in the form of a transfer of the questions into the specific application context. This introduces subjective variability and therefore, makes it difficult to objectively compare the explanatory power between different contexts. However, having a set of template questions can support formulating questions a user might come up with when designing an AI system. This is further supported by works like [27], which investigate the importance of such different questions and criteria. The results of that work highlight that faithfulness and translucence are the most important criteria and that experts and end-users mostly align in how they rate each criterion.

Besides the local or global perspective, measuring explainability can also be categorized into human-centered (e.g. that require human participation via a user study) and methods without human involvement [42]. Approaches without human feedback, for example, because they use an artificial benchmark with defined ground truth for the desired explanations [18, 32, 36], suffer from a direct proof of being transferable to real-world use cases. For example, since explanations are context-dependent, it is not evident that they transfer to other contexts. However, they are significantly cheaper to apply than user studies.

Human-centered evaluations often measure subjective qualities such as user satisfaction or opinion by directly questioning participants. While they provide a holistic perspective on the explanation as they are embedded in the AI system and context, they also introduce additional noise. They do not directly measure the desired qualities but have to rely on proxy questions [42] which introduce human bias [4].

With our study, we want to extend and motivate human-centered evaluations to measure not just satisfaction, but also the degree to which the explanans helps the users to reach their goals in terms of objective task performance.

2.2 User studies on the performance of XAI

Within the health domain, a study on XAI was conducted [13] to evaluate the helpfulness of various approaches to AI assistance in digital pathology. Clinical pathologists were shown various examples of the AI-assistance in a questionnaire, combined with expert interviews. Results of the study show a preference for pathologists for simple visual explanations that correspond to their way of making diagnostical decisions. On the other hand, participants expressed a concern that simplistic explanations allow for a lot of ambiguity in their interpretation.

A study on XAI in the automotive domain [21] investigated the difference in perception of the end users of seeing a textual description of the decision and seeing a textual explanation of the decision without a decision itself. A user was asked to evaluate whether the AI model made a correct decision based on a textual description or an explanation of said decision. Results of this survey show that users tended to trust the explanation without a decision rather than a decision without an explanation.

Lakkaraju and Bastani [23] performed a study on the impact of misleading explanations on users. They hypothesized that existing XAI trust measures are not sufficient, as explanations could be perturbed, leading to users trusting a problematic AI. For this purpose, they constructed a black-box AI making bail decisions based on prohibited features like race or gender. On top, they added an intrinsic explanation method but distorted it such that it returns other, desired features as explanations that were not part of the decision, e.g. prior jail incarcerations. After conducting the study with 41 participants, they could show that users were 9.8 times more likely to trust the black-box AI with a misleading explanation than just the black-box AI although the AI was making the same, wrong decisions in both groups. They advocate for more interactive and explorable explanations, which they back with research findings [24] and a second, slightly adapted study they conducted.

Ribiero et al. [35] present an explanation method that is evaluated in this study, *LIME*. In addition to that, they conduct two studies to evaluate their newly developed method. First, they perform simulated user experiments to investigate the explanation usefulness of *LIME* and compare it to other methods. Second, they evaluate *LIME* with real participants. Here, they checked if users can decide which model performs better, based on the explanations. Ribiero et al. also asked the users whether they trust the decision of a biased model, in this case, the "Husky vs. Wolf" example [3, 19, 33]. They could show, that after displaying an explanation that highlights the model bias, significantly fewer users trusted the bad model.

A different approach presented by [30] is to measure and quantify the amount of information the model provides, by measuring the number of rules a user discovers while interacting with an XAI system. While great at providing concrete

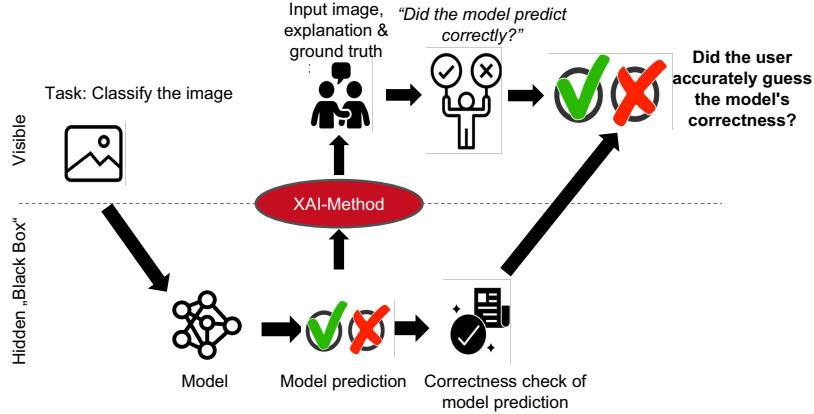


Fig. 1: Overview of the objective methodology for evaluating XAI. It starts with the classification of an image. The model makes a prediction hidden from the user, that will be checked by its correctness in the background. Then, a local post-hoc XAI method is applied to generate an explanation that will be presented to the user, together with the input image and ground truth. Based on this information, the user has to predict the model prediction.

and quantitative numbers the experiment setup, relies heavily on human interpretable features, falling short for complex deep learning models.

A recently published user study is by Achitibat et al. [1] where they explored the practical utility of Concept Relevance Propagation (CRP), a promising development from Layerwise Relevance Propagation. Using provided explanations as guidance, users had to determine if the model's prediction was affected by a bias. In this scenario, two image classification models were used, where one was trained to take advantage of image borders for prediction. For both models, explanations were generated for CRP, but also for other popular XAI methods, e.g. SHAP, Grad-CAM, etc., that were also used in our study. As the design of this user study allowed for objectively true or false replies from the participants, the authors applied common classification result tools such as the confusion matrix and calculated accuracy measures for each method.

3 An objective Methodology for evaluating XAI

A key finding from the literature research was the lack of objective methods for evaluating XAI methods in a human-centered setting. Most of the existing approaches focus on qualitative studies investigating subjective features like satisfaction or just collecting feedback. Quantitative Methods without human involvement are often called to be more objective but are lacking the user focus which is essential for applied XAI methods. The only approach combining human involvement with quantitative methods was by Achitibat et al. [1], where users were asked to assess if the model relied on artifacts in the prediction process.

This approach is a very interesting step in the right direction for quantitative XAI user-studies, but also focused on a sub-task, namely detection of model bias. To allow for a more general approach, our work proposes an objective methodology for human-centered evaluations of XAI methods and executes it on six commonly used explainability methods. An overview of this novel methodology is shown in Figure 1.

3.1 Objective Human-Centered XAI Evaluation

We propose to use a proxy task, to evaluate our approach and answer the research questions about the ability of XAI methods to enable the user to judge, trust, and question AI decisions. The task is designed such that the performance to solve it directly represents the ability to use the explanations for the investigated purpose, e.g. in our case, to determine if the explanation allows the user to judge whether a model predicted the class of an input image correctly. Participants should judge this by reviewing the input image, knowing the ground truth label and the output of a single explainability method. During the experiment, the participants were not aware of the model’s actual prediction or its correctness. An example of this setup can be found in Fig. 1.

Evaluation Metrics As participants judge whether a model’s output is correct, their prediction can either be correct or false. Therefore, their replies can be viewed as a confusion matrix of a binary classifier [45] as summarized in table 1. With this abstraction, it is possible to calculate participants’ accuracy, sensitivity, and specificity. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ expresses the participant’s ability to correctly *judge* model predictions. As we used a balanced number of instances, an accuracy of 0.5 is the baseline for random guessing. This baseline means that in a binary classification task in balanced settings, the simplest random guessing will result in correct judgments about half the time, assuming an equal likelihood of either outcome. This sets a baseline accuracy of 0.5 as the point of comparison, indicating that any performance above this threshold suggests a better-than-random ability to discern the correctness of the model’s outputs. *Sensitivity* and *specificity* are used to measure participants’ ability to identify when they can *trust* a model’s prediction and when they should *question* a model’s prediction. The *sensitivity* is also known as true positive rate $TPR = \frac{TP}{TP+FN}$. The *specificity* is also known as the true negative rate $TNR = \frac{TN}{TN+FP}$.

Table 1: Confusion matrix

	User assumes the model is correct	User assumes the model is wrong
model output was correct	True Positive (TP)	False Negative (FN)
model output was false	False Positive (FP)	True Negative (TN)

Hypotheses testing and effect size To check if the survey results also hold for the general population, hypothesis testing was applied. The significance level α was set to 0.05 for all tests.

A *one-sample t-test* was performed per method. This test is supposed to verify if a sample metric mean of a method can be considered significantly larger or lower than the random baseline. Thus, the null hypotheses (H_0) and alternative hypotheses (H_A) were set according to the resulting sample metric mean size. To measure the effect size of a method in comparison to the random baseline, the effect size *Cohen's d* [8] for the *one-sample t-test* was calculated. The measure is only calculated if the corresponding test result is significant. To evaluate which XAI methods perform best a *paired t-test* was performed for each pairwise method combination.

3.2 Image classification & XAI methods



Fig. 2: Example Survey Question

To avoid model-specific biases, two different AI models, AlexNet [22] and VGG16 [38], were selected for classification. The models were used without hyperparameter optimization. As image source, the *imagenetv2-matched-frequency* dataset [34] was chosen. It is a subset of *imagenet*, which originally contains 1000 classes, and each class has 500 images on average. The objects shown in the images range from everyday items to specific animals.

Based on the experimental setup with a classification of individual images, only local explanation methods fit for the imaging modality could be used. Additionally, the methods had to be applied post-hoc so they could work on the same set of images. Other selection criteria included the use of well-documented methods and the availability of code. Acceptance and relevance within the scientific community were also key factors in the selection process. Following these restrictions, six XAI methods were selected: *Layer-wise relevance propagation (LRP)* [2, 31], *Gradient class activation mapping (GradCAM)* [7, 10], *Local Interpretable Model-Agnostic Explanations (LIME)* [15], *SHapley Additive exPlanations (SHAP)* [28, 29, 37, 40], *Integrated Gradients (Integrated Gradients)* [41] and *Confidence Scores*.

Confidence Scores is the only numerical explanation type used in this study. It provides information on how confident the AI model is that the prediction with first, second, or third rank is correct. It also provides information about the confidence gaps between the ranks. The confidence score is generated based on the normalized results of the last model layer before picking the highest one as the one the model outputs as a prediction. Therefore, it is not a strict probability that the result is correct, but rather a comparative value about how close the highest-ranked decisions were.

3.3 Survey design

Besides a set of demographic questions, incl. educational background, machine learning experience, and visual impairment, the survey consisted of 12 independent questionnaires with 24 pictures each. Of these 24 pictures, 12 were the same across all questionnaires to provide a reliable baseline and the other 12 were semi-randomly picked. They were picked to ensure that each combination of the XAI method, AI model, and output was uniformly represented. Details of this process can be found online³.

To minimize biases, the following precautions were made: All XAI output was used to generate a heatmap which then was overlayed over the original image. Additionally, the heatmaps used the same colormap, one suitable for most color-based visual impairments, and a color bar was added to make the survey as accessible as possible. For reference see Fig. 2.

4 Survey Results

4.1 Questionnaire responses

The survey was closed after one month of execution time. Until the date of 13.07.2022, 139 participants completed the questionnaire. The survey was advertised using the TU Berlin social media pages, resulting in a substantial number of participants with a university background. Of all participants, 24 were undergraduates, while 44 were graduate students. Additionally, 26 postgraduate students took part, as well as seven participants with a PhD. Among the participants, no bias due to education, experience with machine learning, or visual impairment could be identified (see Appendix A for mean accuracy based on demographic groups). A more detailed demographic overview of the participants can be found in Appendix B. To illustrate the results, the *accuracy*, *sensitivity*, and *specificity* results of the 139 participants are shown. The accuracy convergence over an increasing number of participants can be found in the Appendix, see A.2. Figure 3 presents the *accuracy* results. Due to the questionnaire generation procedure, each participant could answer four questions per XAI method. Thus, the only possible results per XAI method were 0.0, 0.25, 0.5, 0.75, and 1.0. The highest mean *accuracy* over all participants was achieved by *Confidence*

³ complete link : <https://github.com/tlabarta/helpfulnessofxai>

Scores with ≈ 0.698 . This was followed by *GradCAM* ≈ 0.603 , *LRP* ≈ 0.583 , *SHAP* ≈ 0.558 , *LIME* ≈ 0.545 and *Integrated Gradients* ≈ 0.532 . Except from *Confidence Scores*, which reached a median of 0.75, all other methods had a median of 0.5. Figure 4a shows two opposite metrics: the *sensitivity*, as well as the *specificity* results. Based on the questionnaire generation procedure, each participant could answer two questions per XAI methods where the models decided *correctly*. Thus, the only possible *sensitivity* results per participant were 0.0, 0.5, and 1.0. The highest mean over all participants was achieved by *GradCAM* with ≈ 0.784 , followed by *Confidence Scores* with ≈ 0.752 , *LRP* ≈ 0.748 , *LIME* ≈ 0.590 , *Integrated Gradients* ≈ 0.561 and *SHAP* ≈ 0.335 . *GradCAM*, *Confidence Scores* and *LRP* reached a median of 1.0, *LIME* and *Integrated Gradients* a median of 0.5, whereas *SHAP* reached a median of 0.0. Due to the chosen questionnaire generation procedure, each participant could answer two questions per XAI method where the chosen models decided *incorrectly*. Thus, the only potential *specificity* values per participant were 0.0, 0.5, and 1.0. The highest mean over all participants was achieved by *SHAP* with ≈ 0.781 , followed by *Confidence Scores* with ≈ 0.644 . *Integrated Gradients* reached ≈ 0.504 , *LIME* = 0.5, *GradCAM* ≈ 0.421 and *LRP* ≈ 0.417 . *SHAP* reached a median of 1.0 whereas the other XAI methods had a median of 0.5. Participant responses could be influenced by the model that an explanation was generated for.

As described in section 3.3, the predictions were split half/half between both models. Since VGG16 generally achieves a better task performance than AlexNet one would assume a noticeable impact on the participant accuracy across all XAI methods. The assumption was that a better-performing model leads to more profound decisions and weights, which would have a positive impact on the generated explanations. Figure A.3 in the Appendix shows that only a small difference between the two models existed. Participants performed marginally better when explanations were generated on VGG16 predictions, with an accuracy of 0.61 versus 0.56 for explanations generated on AlexNet predictions. There is little to no difference in participant performance between the two models for *Integrated Gradients*, *SHAP* and *GradCAM*. A noticeable difference can be seen for *Confidence Scores*, *LIME*, and *LRP*.

4.2 Qualitative Feedback

In addition to the questionnaire responses, participant feedback about the study was received. Some of the feedback could be valuable for future studies. One of the participants pointed out that it was hard to decide without knowing if a model was trained to recognize a specific class. On one hand, it would help the participant to make a decision, but on the other hand, it would also enable certain bias since it is possible to assume a participant would rather tend to answer "yes" to those classes which are in the trained classes list, and rather "no" to those classes not on the list. Also due to the broad participant scope of this study, it can be assumed that this information would not be helpful for all participants. Another point was in the case of multiple objects shown in a picture. Feedback was that it would be helpful to highlight the object to be classified if

there is more than one. For example, when having multiple dog breeds on an image, it would be good to know which one is to be classified. A disadvantage of this approach is that the model itself is not "told" which object it is asked to identify, this task belongs to the challenge of image recognition as well. Another feedback was, that even people with a machine learning background had in some cases a hard time making a decision. This could explain the observed very low-performance differences between machine learning experts and non-experts.

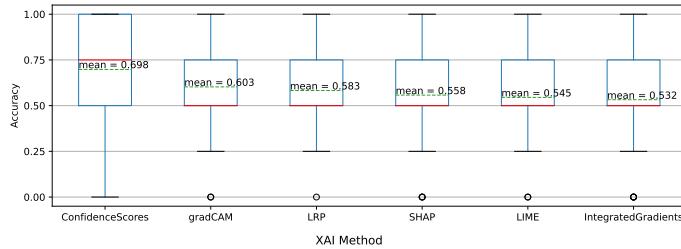


Fig. 3: The mean accuracy of participants in determining the correctness of the AI's predictions varied by XAI method used. Confidence Scores consistently outperformed all other methods, while the remaining methods showed minimal to no differences in results.

5 Discussion

Continuing with the discussion, table 2 states the results for the *one-sample t-test* performed on the accuracy means. All means were significantly greater than the random baseline of 0.5 (better than chance). Only *Confidence Scores*'s accuracy was significantly higher than for all other methods. These findings were drawn from the pairwise conducted *paired t-tests*. Thus, *Confidence Scores* helped a user most to *judge* an AI's decision with a strong effect. All other methods were similarly helpful, with small to no practically measurable effect.

It can be assumed that *Confidence Scores*'s superior *accuracy* (i.e. overall performance) originates partially from its relatively simple structure and clear results of just class probabilities. The explanations generated by the other methods have considerably more information than just classes with a probability and therefore need more interpretation. Participants had to consider the location of highlighted areas, as well as the color and brightness.

Furthermore, table 3 states the results for the *one-sample t-test* performed on the *sensitivity* and *specificity* means. All XAI methods but *SHAP* had means that were significantly greater than 0.5. However, only *GradCAM* had a large positive effect.

Overall, *GradCAM*, *Confidence Scores* and *LRP* helped a user most to *trust* an AI's decision with a medium to strong effect. Interestingly, *SHAP* even negatively impacted trust in an AI's decision with a medium negative effect. *SHAP* and *Confidence Scores* achieved means that were significantly greater than 0.5. *Integrated Gradients*'s and *LIME*'s means were not significantly different from

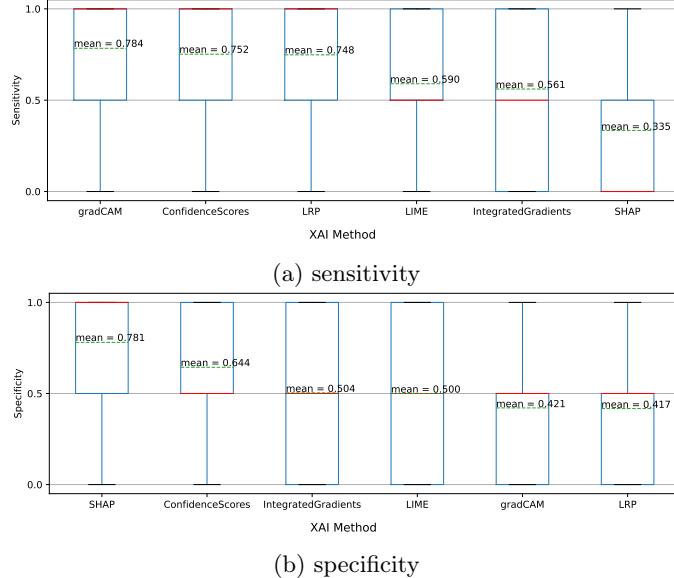


Fig. 4: Participants' mean sensitivity & specificity in judging whether the AI was correct in its prediction, per XAI method. SHAP performs best for sensitivity and worst of all methods for specificity. Confidence Scores perform the best in both measures, besides SHAP.

Table 2: Hypothesis testing: accuracy

XAI method	Accuracy	HA	p	d
<i>Confidence Scores</i>	≈ 0.698	> 0.5	$< .001$	0.84
<i>GradCAM</i>	≈ 0.603	> 0.5	$< .001$	0.46
<i>LRP</i>	≈ 0.583	> 0.5	$< .001$	0.40
<i>SHAP</i>	≈ 0.558	> 0.5	$< .001$	0.28
<i>LIME</i>	≈ 0.545	> 0.5	.009	0.20
<i>Integrated Gradients</i>	≈ 0.532	> 0.5	.04	0.15

0.5. *GradCAM* and *LRP* means were significantly smaller than 0.5. Thus, *SHAP* helped a user most to *question* an AI's decision with a strong effect.

The Imagenet dataset was chosen because of the assumption, that most people naturally are domain experts in classifying its images, as it shows objects of daily life. However, some participants mentioned that they had never seen some of the fruits, animals, or other objects that were presented before, meaning they lacked the required domain knowledge. This naturally impacts the ability to interpret the explanation, such as not being aware of the anatomic unique features of certain dog breeds.

Usually, XAI methods are applied to help domain experts in evaluating model predictions. These should be aware of relevant details within their domain and their response might vary compared to our survey. In most applications, the field

Table 3: Hypothesis testing: specificity & sensitivity

XAI method	Specificity	HA	p	d	Sensitivity	HA	p	d
<i>SHAP</i>	≈ 0.781	$> 0.5 < .001$	0.85		≈ 0.335	$< 0.5 < .001$	-0.43	
<i>Confidence Scores</i>	≈ 0.644	$> 0.5 < .001$	0.44		≈ 0.752	$> 0.5 < .001$	0.77	
<i>Integrated Gradients</i>	≈ 0.504	> 0.5	.458	not significant	≈ 0.561	> 0.5	.037	0.15
<i>LIME</i>	0.5	$\neq 0.5$	1.0	not significant	≈ 0.590	> 0.5	.004	0.23
<i>GradCAM</i>	≈ 0.421	< 0.5	.007	-0.21	≈ 0.784	$> 0.5 < .001$	0.89	
<i>LRP</i>	≈ 0.417	< 0.5	.002	-0.25	≈ 0.748	$> 0.5 < .001$	0.74	

of predictions is much leaner and more focused on a specific application than the 1000 classes presented in ImageNetV2.

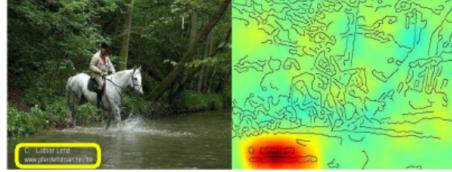


Fig. 5: Right prediction based on a wrong feature

Another critical point is the potential effect of so-called model bias on the survey results. It occurs in cases when a model makes a correct prediction based on wrong features. An example of this phenomenon can be seen in Fig. 5 where a horse picture from Pascal VOC data set was classified correctly by Fisher vector classifier [25] because of a photographer tag in the bottom left corner. By seeing the explanation of this example, it should be possible for most users to detect the model bias and conclude that the model did not classify this image correctly. Detecting such model bias usually requires expert knowledge in the machine learning field.

Noteworthy is also a potential bias in the study population, as it was shared within university context for everyone to attend, without demographic selection criteria. When analyzing participant performance based on demographic criteria, no statistically significant impact could be identified, however, that does not exclude potential bias from other demographic criteria that were not tracked.

Since a model made the right classification, even though based on completely irrelevant features, user decisions would contribute to false-negative samples in a confusion matrix. In the context of this survey, it reduces the accuracy and sensitivity of the XAI method. The problem here is that even though an XAI method showed a good explanation of the decision, its performance would still be degraded since the model did a bad job of learning the class features. This is a potential issue that could not be fully excluded from the survey, as there was no documented list of identified model biases available for VGG16 or AlexNet on ImageNet.

To summarize, the results showed a clear superiority of the heuristic explanation method *Confidence Scores* in *accuracy* and a good performance in *sensitivity* and *specificity*. A potential risk of the method is that purely heuristic explanations

can lead to heuristic biases as they tend to oversimplify complex situations [44]. This oversimplification and an high enough trust on the users side can lead to over reliance [11], in which the users trust the model to much and let it cloud their judgement. A good explanation method needs to account for this by clearly visualizing the relevant features and combinations, to make it as easy as possible for the user to identify wrong predictions. In terms of *specificity*, *SHAP* was significantly above random baseline and superior to all other methods although it performed very badly for *sensitivity* and slightly above random baseline for *accuracy*. All in all, no explanation method overperformed in all cases.

The results indicate that it would therefore be recommended to not rely on a single, monolithic explanation approach. Instead, a more diverse and interactive framework of explanation methods could be beneficial. This explanation framework should be focused on the explanation task and the user's needs. Diverse in this context is understood as a selection of different explanation approaches rather than a single method. Such a framework could also help with reducing the previously discussed issues of the required domain and expert knowledge for certain explanation cases and could be more effective in detecting model bias. The design of such an explanation framework, as well as its evaluation, would be a research topic for the future.

A key contribution of this work was the proposal of a new, objective methodology for evaluating XAI. The survey executed with this methodology shows the benefit of the approach, extending highly subjective aspects of human-centered XAI evaluation with task performance-related ones. A point of critique is the relatively staged scenario of users not knowing the model decision. Usually, users should be aware of the decision but might not be able to judge whether it is correct or not. A thorough examination of the objective methodology as well as an approach that is closer to reality should also be part of future research.

6 Conclusion

Within this paper, an objective methodology for evaluating XAI methods was proposed. From literature research, a lack of quantitative methods for human-centered XAI evaluation was identified, which this work aimed to contribute to. This approach was evaluated in a user study, where six state-of-the-art XAI methods were implemented. The goal of the user study was to examine how far existing methods enable users to judge, trust, and question AI decisions.

The results show that of the tested methods only *Confidence Scores* substantially enabled users to judge an AI decision, responding to the first research question. Aimed at the second research question, the methods *GradCAM*, *Confidence Scores* and *LRP* performed best in making users trust an AI decision. For the final research question, it can be concluded that *SHAP* enabled users by far the most in questioning an AI decision.

From the research findings but also literature research, it can be concluded that using individual explanation methods is not sufficient for enabling users to judge, trust, and question an AI effectively. Instead, the design of an interactive

framework of multiple explanation methods was proposed, to achieve better user focus. Further, we would suggest setting up studies that do not just measure the user’s classification performance on detecting true AI decisions but also measure if they can detect non-wanted biases or shortcuts [16]. A shortcoming of the proposed evaluation method is that it is still relatively costly to execute, being a user study requiring human involvement. We hope to see further human-centered XAI studies that extend our approach to further proxy tasks to create a solid ground for future XAI research and the creation of trustworthy AI applications.

Disclosure of Interests The authors declare no conflict of interest.

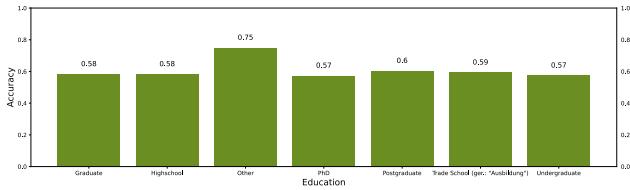
References

1. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence* **5**(9), 1006–1019 (2023)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* **10**(7), e0130140 (2015)
3. Badillo, S., Banfai, B., Birzele, F., Davydov, I.I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., Zhang, J.D.: An introduction to machine learning. *Clinical pharmacology & therapeutics* **107**(4), 871–885 (2020)
4. Bertrand, A., Belloum, R., Eagan, J.R., Maxwell, W.: How cognitive biases affect xai-assisted decision-making: A systematic review. In: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society. pp. 78–91 (2022)
5. Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., Holzinger, A.: Quod erat demonstrandum?—towards a typology of the concept of explanation for the design of explainable ai. *Expert Systems with Applications* **213**, 118888 (2023)
6. Carli, R., Najjar, A., Calvaresi, D.: Risk and exposure of xai in persuasion and argumentation: The case of manipulation. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 204–220. Springer (2022)
7. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)
8. Cohen, J.: Statistical power analysis for the behavioral sciences. Routledge (2013)
9. Council of European Union: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance) (May 2016), <https://gdpr.eu>
10. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020)
11. Du, Y., Antoniadi, A.M., McNestry, C., McAuliffe, F.M., Mooney, C.: The role of xai in advice-taking from a clinical decision support system: A comparative user study of feature contribution-based and example-based explanations. *Applied Sciences* **12**(20), 10323 (2022)
12. of European Union, C.: Ethics guidelines for trustworthy ai. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>, accessed: 2022-12-30
13. Evans, T., Retzlaff, C.O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.R., Zerbe, N., Holzinger, A.: The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems* **133**, 281–296 (2022)
14. Farhat, H., Sakr, G.E., Kilany, R.: Deep learning applications in pulmonary medical imaging: recent updates and insights on covid-19. *Machine vision and applications* **31**(6), 1–42 (2020)
15. Garreau, D., Mardaoui, D.: What does lime really see in images? In: International Conference on Machine Learning. pp. 3620–3629. PMLR (2021)

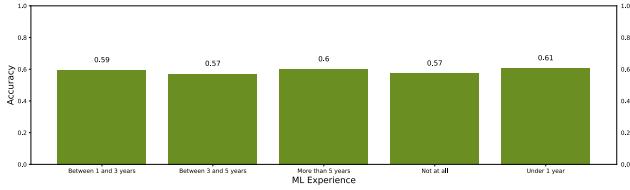
16. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
17. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* **37**(3), 362–386 (2020)
18. Hedström, A., Weber, L., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: an explainable ai toolkit for responsible evaluation of neural network explanations. *arXiv preprint arXiv:2202.06861* (2022)
19. Hodges, J., Mohan, S.: Machine learning in gifted education: A demonstration using neural networks. *Gifted Child Quarterly* **63**(4), 243–252 (2019)
20. Hu, X., Chu, L., Pei, J., Liu, W., Bian, J.: Model complexity of deep learning: A survey. *Knowledge and Information Systems* **63**(10), 2585–2619 (2021)
21. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Proceedings of the European conference on computer vision (ECCV). pp. 563–578 (2018)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60**(6), 84–90 (2017)
23. Lakkaraju, H., Bastani, O.: " how do i fool you?" manipulating user trust via misleading black box explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 79–85 (2020)
24. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and customizable explanations of black box models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 131–138 (2019)
25. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. *Nature communications* **10**(1), 1–8 (2019)
26. Liao, Q.V., Gruen, D., Miller, S.: Questioning the ai: informing design practices for explainable ai user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2020)
27. Liao, Q.V., Zhang, Y., Luss, R., Doshi-Velez, F., Dhurandhar, A.: Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 10, pp. 147–159 (2022)
28. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* **2**(1), 56–67 (2020)
29. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017)
30. Matarese, M., Rea, F., Sciutti, A.: How much informative is your xai? a decision-making assessment task to objectively measure the goodness of explanations. *arXiv preprint arXiv:2312.04379* (2023)
31. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* pp. 193–209 (2019)
32. Müller, H., Holzinger, A.: Kandinsky patterns. *Artificial Intelligence* **300**, 103546 (2021)
33. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* **116**(44), 22071–22080 (2019)

34. Recht, B., Schmidt, L., Roelofs, R., Shankar, V.: Imagenetv2. <https://Imagenetv2.org>, accessed: 2022-09-17
35. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
36. Salewski, L., Koepke, A., Lensch, H., Akata, Z.: Clevr-x: A visual reasoning dataset for natural language explanations. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. pp. 69–88. Springer (2022)
37. Shapley, L.S.: A value for n-person games. Classics in game theory **69** (1997)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Sovrano, F., Vitali, F.: How to quantify the degree of explainability: Experiments and practical implications. In: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–9. IEEE (2022)
40. Speith, T.: A review of taxonomies of explainable artificial intelligence (xai) methods. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 2239–2250 (2022)
41. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
42. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion **76**, 89–106 (2021)
43. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating xai: A comparison of rule-based and example-based explanations. Artificial Intelligence **291**, 103404 (2021)
44. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI conference on human factors in computing systems (2019)
45. Yerushalmy, J.: Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. Public Health Reports (1896-1970) pp. 1432–1449 (1947)

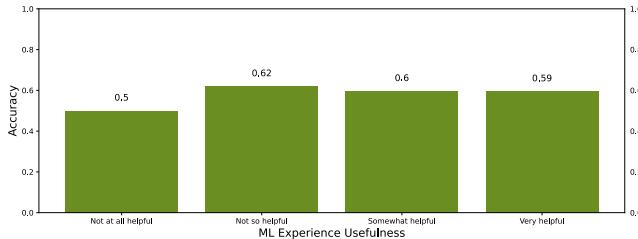
Appendix A Additional Visualizations



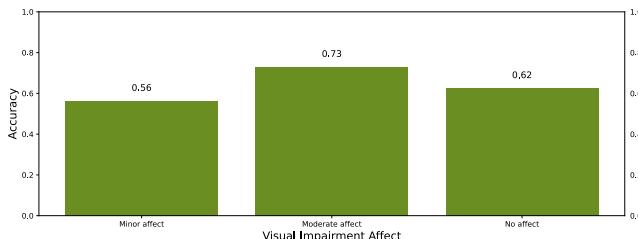
(a) educational background



(b) machine learning experience in years



(c) self-assessment of the usefulness of machine learning experience for answering the survey



(d) self-assessment on visual impairment

Fig. A.1: Mean Accuracy based on educational background, machine learning experience in years, self-assessment of the usefulness of machine learning experience, and self-assessment on visual impairment.

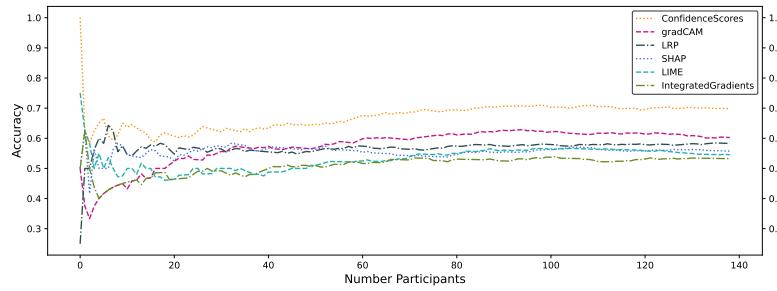


Fig. A.2: Convergence of accuracy over the number of participants.

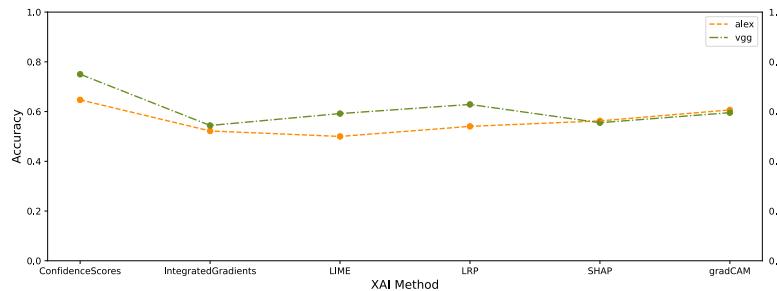


Fig. A.3: Difference in accuracy between VGG16 and AlexNet

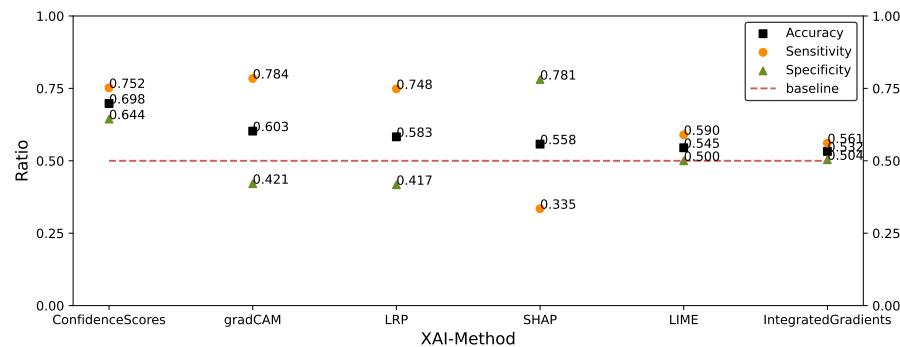


Fig. A.4: Summary of means for accuracy, sensitivity, and specificity for all examined XAI methods compared to the random baseline.

Appendix B Demographic overview of participants

Table 1: Results of demographic questions

Education								
High School	Trade School	Undergraduate	Graduate	Post Graduate	PhD	Other		
20	4	24	44	26	7	1		
ML Experience								
None	Less than 1 year	Between 1 and 3	Between 3 and 5	More than 5				
65	33	26	7	5				
Perceived helpfulness of ML experience								
Not at all helpful	Not so helpful	Somewhat helpful	Very helpful					
2	15	43	11					