Investigation of the evaluation techniques and tools used for model-specific XAI models

Tanguy Marbot
Supervisor(s): Chhagan Lal, Mauro Conti
EEMCS, Delft University of Technology, The Netherlands

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
20-6-2022

Abstract

The spread of AI techniques has lead to its presence in critical situations, with increasing performance that can compromise on its understanding. Users with no prior AI knowledge rely on these techniques such as doctors or recruiters with a need for transparency and comprehensibility of the mechanisms. The advent of Explainable Artificial Intelligence techniques responds to such issues with a diversity that has lead to the construction of a taxonomy for this domain. Notably, there is a distinction between model-specific and model-agnostic techniques. Rightly operational XAI technique should go through an evaluation process. In this paper, we investigate the different available tools and metrics for the evaluation of XAI techniques to then assess the evaluation quality of five state-of-the-art model specific techniques: TCAV, SIDU, ACE, Net2Vec and Concept Analysis with ILP. It has been concluded that despite broad existing literature on evaluation methods, there is a lack of exhaustive assessment of criteria and a lack of standardization in regards of the evaluation of these model-specific-techniques.

1 Introduction

Within Artificial Intelligence (AI), the exponential rise of different models has opened many opportunities for assistance and automation. The growth of data resources comes hand in hand with the omnipresence of AI systems, for example in self-automated driving, in the medical sector and media [12] [5] [20].

A great progress in performance is paralleled with an urgent need to address issues such as transparency, trust and accountability in case of harmful impact. Indeed, AI systems might reflect unfair biases from data sets, for example in assisted job interviews. Additionally, they might need to make critical decisions in case of car accidents [5]. In general, the black-box nature of AI models inhibits the access to crucial information [20].

Explainable Artificial Intelligence (XAI) techniques have been developed to tackle such issues [21] [2] [19], attracting interest in research. The field of XAI is large with different approaches with different intents. For example, there can be a trade-off between the correctness of information and the interpretability (the ability for a human to understand the information) [20]. The choice of which of the two to favor can depend on the level of AI expertise of the user.

A taxonomy of the domain has been developed [5], specifying different types of XAI techniques. One of them distinguishes methods under the interpretability technique with two fundamental categories: model-agnostic techniques and model-specific techniques. The first type applies when the technique could be used generally. The second category is for techniques that are designed in focus of an AI technique in particular, [20]. Due to a diversity of techniques and goals falling under both categories, it is worthwhile to focus research to one domain and evaluate the efficiency of different techniques in respect of intended purposes.

Moreover, the field has developed with diverse application domains, and explanations can be more specific for the sake of better performance. Therefore, one might need to choose between certain trade-offs depending on personal knowledge or the desired performance of a surrogate model for explanation. Surveyed evaluation criteria and metrics can be a useful resource to guide experts in the process of techniques design and improvement as well as guidance for non-expert users to choose a suitable technique.

Current literature has already examples of surveys on the evaluation methods and metrics. However, there seems to be a gap in the research space regarding an investigation on the evaluation of model-specific techniques.

This paper aims to explore and explain the means of evaluation of efficiency and correctness for model-specific papers. The first contribution is to make an investigation of the evaluation techniques of XAI in general, second contribution is to compare the evaluation process of state-of-the-art model-specific XAI techniques and highlight the similar and neglected tools of evaluation. The following question is thus answered "How are the proposed state-of-the art model-specific techniques evaluated to show their efficiency and correctness ?".

Section 2 of this paper presents the background literature used and the approach of the research. We will first investigate the evaluation tools and metrics already enunciated in the current literature, to build an exhaustive list of explanation criteria and tools in Section 3. Then, in Section 4, we will compare the evaluation process of five model-specific techniques. Section 6 is a discussion about our findings with a debate on the relations between the third and fourth section. The final section suggests future work directions.

2 Background and Related Work

Our methodology will be the one of a literature review with particular requirements from the background literature.

To assess explanation techniques, there is first a need to rely on the clear formulation of criteria that correspond to explanation goals of a technique. As the notion of explanaibility is subjective and due to a diverse nature of contexts of explanation depending on end users and the issue in question, we might favor certain aspects of explanation over others [4]. For example, there would be a difference between one explanation for AI experts, and one for users who are not AI experts. Doctors using a diagnosis AI assistance, need to benefit from explanations to work on a specific medical task, whereas AI experts such as designers might benefit from explanation to perform debugging and improve the system. Thus, firstly, metrics and criteria need to be investigated and characterized with regards of the contribution to explanation objectives. Indeed, we need to find why a particular metric is important for the evaluation (i.e., how it contributes in the evaluation process and to what extent). Then, we can find tools used for such measurements and check if there are any formal or theoretical approaches for evaluations.

The literature already provides some answers to these ques-

Technique	Summary
TCAV [14]	Uses linear classification in any layer, and directional derivatives to achieve the quantification of the classification sensitivity of a concept given by a user through example samples.
SIDU [18]	Localizes entire entire object regions responsible for prediction. Applies to CNN models by using convolutional layer and mask generation.
ACE [1]	An automated use of TCAV. Takes images of same class as input and a trained CNN to build image segments and cluster them together. TCAV will compute the importance score of the segments
Net2Vec [10]	Makes use of combination of filters which responses can construct vectorial embeddings to which semantic concepts can be mapped to.
Concept Analysis with ILP [13]	Derives symbolic knowledge in the inner layers of a DNN model and uses an ILP model to build explanation in the form of first-order rules

Table 1: Five model-specific XAI methods

tions [17] [20] [4] [8]. The main sources used in this paper offer a wide range of evaluation tools and criteria, but sometimes seem to lack consistency between each other, as some terms and meanings are used interchangeably with different implications, for example,, "stability" and "robustness", "faithfulness" and "fidelity". Moreover, because criteria can be homonymous with specific metrics of measurement, in this paper we will refer to the term "metric" as both the specific measurement tools (e.g Area Under Curve) and the desiderata which measurements respond to (e.g. fidelity).

Also, papers don't always provide examples of actual quantifiable measurements used for all metrics. Our paper attempts to address these issues of consistency and exhaustivity by analysing, more specific papers, for example related to a single criteria, such as in [3] or [22] as well as the different existing surveys for comparison.

After the investigation that provides a general view on the process of evaluation, we can collect state-of-the art model-specific techniques papers and assess, with regards of the first investigation part, each evaluation process. This assessment leads to extract observations about the neglected tools and other note-worthy information for discussion. Table 1. presents a summary of the five chosen techniques.

3 XAI evaluation concepts

We are categorizing criteria and their metrics under the taxonomy of functionally grounded metrics, Human grounded metrics and Application grounded metrics. Table 2. presents a summary of the three main types of metrics. It is important to note that other forms of distinction exist. For example, distinction can depend on the nature of explanation, [4]. The explanation can stem from self-explanatory or surrogate models, which we call model-based explanation. In case of attribution of existing elements of the initial model, we can refer to attribute-based explanation. If we use examples as explanation, in case for example of counter-factual, we have example-based explanations.

3.1 Functionally grounded metrics

Functionally grounded metrics use proxy based on properties of explainability that can be directly measured by computation. Criteria tend to be focused on the correctness of the explanation, in contrast with interpretability, which is not as evident to automatically quantify. The evaluation of functionally grounded metrics is usually less costly than the other types of metrics.

Fidelity

An XAI technique is essentially an AI technique with its perceivable information being substituted or supplemented. The substituted or additional information can be more or less in accordance with the initial material. It is primordial that an explanation reflects reality, as there would be no use of an explanation that is misleading. For example, in an image classification network, if an explanation highlights the importance of certain regions of an image for its class prediction, it should be the case that this region was important in the actual class prediction to make the explanation sound and useful.

This congruity between explanations and reality is called **fidelity**. In other words, fidelity can be designated as the requirement of objects of explanations being fully incorporated into the decision-making process. More generally, it is referred to as the accuracy of the explainer regarding the object of explanation.

Swartout and Moore (1993) proposed that fidelity, understandability, and sufficiency were properties of good explanations [1]. This is reflected in the literature as it is generally part of the first metrics to be measured, sometimes the only metric to be measured, [13].

Another synonymous term for fidelity is **faithfulness**. The definition of fidelity and faithfulness can differ, and the terms can be used interchangeably in the literature, or with a clear distinction in [20] . Here, faithfulness can refer to feature importance related explanations and fidelity to surrogate models, which is the case of the most popular model-agnostic techniques, for example LIME.

High fidelity might come with some trade-offs. It can come into discussion when comparing inherently explainable models and post-hoc explanations when a new model is created to provide explanations to the first one. The latter one can have better performances, leaving the initial model intact, but might compromise on accuracy. The first one has undistorted, fully faithful explanation but can sacrifice on performance, [9]. There is also a trade-off between interpretability and faithfulness as a simpler explanation will omit certain cases.

Fidelity is usually computed automatically. Since humans tend to prefer simplified explanations, relying on a human-grounded evaluation of explanation could lead to a lack of transparency and performance for the sake of persuasiveness.

	Task	Subject	Cost	Evaluation metrics
Functionally-grounded	Proxy	Automated	Lower	Fidelity, Robustness, Correctness, Safety, Architectural complexity, Expressiveness
Human-grounded	Proxy	Humans	Higher	Simulatability, Trust, Preference, Comprehensibility, Time Efficiency, Amount of information, Debuggability, Model Validation, Time Efficiency
Application-grounded	Application Interactions	Humans	Highest	Performance, Satisfaction, Persuasiveness, Human Judgement, Novelty

Table 2: Taxonomy of the evaluation process with three categories: functionally, human and application grounded metrics. The task refers to what is being directly assessed, Proxy meaning a mediate task that can assess a specific property of the technique. The subject refers to the agent for the task.

[17]. Therefore, it is preferable to use computed evaluation of fidelity.

In general, metrics for the measurement of fidelity are quantifiable and can rely on using as signal the model behavior change. A case for evaluation are saliency methods. Saliency methods highlight, by assigning scores, features that are deemed as the most relevant in a classification. To verify whether the scores reflect true importance, an approach could be to monitor the model behaviors after strategical modification of the input according to the explanations. IAUC, area under the insertion curve, is a possible metric. Starting with a reference input, constant value or blurred image, we insert features in the order of relevance (from high to low) and measure the probability increase. The higher the surge at the beginning, the higher the IAUC and thus the fidelity, [15].

Authors in [7] show a model-specific technique where we can form clusters from data input instances, that essentially derive explanation from the inner learning process of random forests and evaluate the accuracy of instances inside clusters with actual results.

Fidelity is often associated with **completeness** or **coverage**. They serves to measure the size of a high-fidelity subset, how large is the range of validity. In other words, they measure the scope of the explanation, whether the XAI technique managed to highlight relevant features exhaustively. They can be seen as a generalization of fidelity. Another metrics associated with fidelity is **localization accuracy** which can be measured with the **PG** method, in the example of saliency maps, by checking if highest score pixel lies in the ground-truth area. [15].

Robustness

Robustness refers to the stability of the output, when small changes are made in the input. In other words, a model is robust when similar inputs result in similar outputs, [20]. The robustness of explanation shouldn't be confused with robustness of the prediction model which can assessed with safety metrics, [4].

Concepts related to robustness can be identity, stability and separability. Identity ensures that identical instances give identical results whereas stability ensures that similar instances give similar results. Lastly, separability asserts that disparate explanations need to arise from disparate instances, [11]. We will group them in this paper under the term robustness

Robustness complements fidelity in terms of explanation quality. David Melis and Tommi Jaakola argue for the importance of assessing robustness for saliency methods that are not invariant and are sensitive to the choice of reference point. Also, it might be too optimistic to understand a complex model with limited scope explanation on single points. Thus, robustness contributes to a rigorous evaluation of the explanation, [3]. Moreover, robustness ensures the resilience of the model against attacks, e. g. adversarial attacks that make non-perceivable changes to input, [18].

Assessing the robustness relies on making slight change to the input of the model, such that the prediction doesn't change. We can then observe the discrepancy of the explanation using calculus tools such as the notion of local Lipschitz continuity, [3].

Correctness or accuracy

Another quantifiable performance related metrics is **accuracy**, which relates to the ability of the explanation model to make correct predictions independently from the object of explanation. This criterion can be fused to the **correctness** of the explanation model. Not only this is required to ensure reliability and the trust of the user, but authors in [22] have shown that classification accuracy is positively correlated with explanation accuracy.

To measure accuracy one can for example simply take the inverse of the weighted sum of the boolean values of the equality between the prediction of the explanation model and the true prediction, for all input. Authors in [16] refer to this metric as **Classification Accuracy**. Furthermore, authors in [22] (Section 2.2) Show a metric called **k-accuracy** in the context of strings in the dataset.

Safety

Also called privacy. Measures the exposure of critical information. An explanation shouldn't unveil sensitive information to anyone. No clear metrics for measurement could be found in the literature, [20].

Architectural complexity

Architectural complexity quantifiably addresses the interpretability of the system, i.e. how understandable is the explanation for a user, as opposed to a focus on correctness of explanation. This criterion can be assessed by directly measurable properties of the explanator type, [8]. This often relates to size measures which goal is to approximate per-

ceived human-grounded complexity and favor more reliable and compact model explanation.

Metrics can for example be the sparsity of linear models, number of used input features, etc. In the case of decision tree, it can include the depth of tree or the number/length of rules. Some definition of explanability can be tied to simulatability, the ability for the user to make predictions when given solely input and explanation of the model. With this assumption, we can measure complexity based on the number of runtime operation counts, the number of arithmetic and Boolean operations performed by the explainable model when given an input by the target of explanation, [4].

This criterion can be further expanded by **algorithmic complexity**, which relates to the theoretical complexity of the algorithm used to produce the explanation, this refers to the time to convergence to a solution, [20].

Expressiveness

Expressiveness is another interpretability related computable criteria. We can measure the amount and density of information perceived by the user and the level of details provided by the explanator, a detailed explanation increases the chance to understand what is going on.

There are different metrics at disposal:

-The number of expressible relations

-The depth of added information or the measure of information units used per explanation

-The type of expressiveness of used rules. For example, boolean or first-order logic, [20].

3.2 Human grounded metrics

Human grounded metrics is about involving human-subject, experiments on proxy tasks for measurement. Domain experts are not required, and the target application is evaluated in its essence even though the final application is not run to avoid the need of experts and to save time. [14] [20]. Such metrics are often used to measure more general concepts of explainability, psychological ones for example the quality of the mental model of the user.

Furthermore, a good explanation should confirm certain desiderata of the AI model:

-Privacy as sensible in the data shouldn't be exposed.

-Fairness should protect groups from discrimination and unjustified biases. [6]

Simulatability or predictability

One evidence for the understanding of the user is his capacity to simulate the behaviour of the explainable model. As an experiment after explanation, humans could be presented with an explanation and an input. Then, they are expected to correctly determine the model's output without knowledge of the initial input of the given explanation, The assessment of the accuracy of the user's simulation can then be computed using proxies of fidelity. [20]

A variant is counterfactual simulatability, [8] After being presented with an input, an output and an explanation, they are presented a different desired output and are asked what modification of the input must be done to change the method's prediction to the desired output.

Trust and Preference

The model explanation can be assessed by the appreciation of the user through different criteria such as **confidence** or **trust** and **preference** which are primordial for evaluation of XAI, [4]. A good explanation should be preferred and trusted by the user. Confidence and trust can be used interchangeably, we will refer both as **trust**, the trust a user has on a given explanation.

A rather simple way to assess preference is through a binary choice. The user must choose one preferred explanation from a pair presented to him, [8]. In general, subjective questionnaires can be designed for users who can be asked after or during task time to get subjective responses, [4]. It is noticed for example that explanation of training data points influence has a considerable effect on user trust. Those questionnaires can also keep track of the degree of understanding over time, where there can be different status of understanding characterized in the process, which helps finding the necessary measures to complete the mental model. [20]

Comprehensibility

We need to be sure that the user can integrate and make sense of the information provided.

Comprehensibility assessment usually relies on subjective feedback and depends on the background of the user (vocabulary, biases, ...). It is also possible to extract information in an objective way, we can measure human metrics, such as behavior and physiological signs. For example, the Blood Volume Pulse (BVP) and Galvanic Skin Response (GSR) are influenced by the explanation presentation. We can also quantify response times and accurate decisions in decision making processes, where both can indicate a more intuitive understanding of explanation. [4]. Moreover, we can assess the quality of the mental model with the length of the user's self-explanation. [17]

Amount of information

Analogous to architecture complexity but assessed subjectively by a user, [20]. An excessive amount of information might impede on the **interpretability**.

Debuggability and model validation

Focused on a developer use case, **debuggability and model validation** refers to the leverage that explanation provides to improve the model. There is thus a clear notion of utility of the explanation in contrast of trust, as we are not measuring a subjective form of validation of the system, but we are assessing the potential contribution, by the explanation, to further advancements. This is especially helpful for designers. [4]

Downstream tasks (supervised tasks given to the user benefiting from explanations) are possible to assess such utility. For example:

- Given an incorrect model decision and corresponding explanation, determine the reason the model made a mistake. [6]

Time Efficiency

Time efficiency measures the time it takes for the user to construct a viable mental model, as a an explanation shouldn't take too long to be understood. For measurement, we could

use tools for the assessment of **simulatability** tracked with time.

This efficiency is critical in situations such as automated driving applications or recommendations systems, where the explanation needs to be assimilated in a limited time frame. [20]

3.3 Application-grounded metrics

Application Grounded Metrics are metrics that can be measured in the context of the interaction of the user with the intended final application after benefit of explanation. For example, one might be interested in the performance of doctors performing diagnoses, where assistance of an XAI is provided. [20] [8]

Performance of User-AI System

We can see the user and the AI system as members of a same team than can improve its performance thanks to easier collaboration made possible with explanation. This can apply, for example, for a doctor with a medical assistant system who may benefit from explanations to correct the assistant in case of errors and thus prevent wrong diagnoses, [20] [6].

We can directly measure the performances of the end-to-end task while comparing two situations: one with the user-AI assisted system with explanation and one without explanation. Therefore, measurement of **performance** difference directly attests the utility of explanation.

Specific metrics can be derived from the measure of performance: ability to detect errors, accuracy, response time needed, likelihood to deviate [4].

Satisfaction

We can implicitly measure the quality of explanation by questioning the satisfaction judgement of the user of the application with explanation provided. This is can be done with questionnaires, [4] [20].

Persuasiveness

Measures the ability of the explanation to push the user in a certain direction. This is particularly useful in recommending systems, where the goal is to direct the users toward certain choices, [20].

Human judgement

Measures the level of appropriate trust in the decisions of the system. A healthy human-machine implies that a user should reject incorrect decisions proposed by the model and accept correct decisions. [20] Brittany Davis, Maria Glenski, William Sealy and Dustin Arendt [6] Propose a downstream task where, given a series of inputs, users can make decisions of agreement or disagreement with the model's output. The results are compared between a situation with explanation and one without.

Novelty

An explainee can perform better when the tasks are less repetitive and when he feels less bored. This is achievable by a subjective degree of information novelty, which can be related to satisfaction and efficiency [20]

4 Comparison of model-specific techniques evaluation

We will now conduct an analysis of the evaluation of five model-specific XAI techniques in regards of the metrics we have just found earlier. This will provide us a basis of comparison and underline eventual neglects. We make a description of each method's evaluation process and Table 3 provides a comparison between the techniques evaluations with the functionally, human and application-grounded taxonomy.

4.1 TCAV

The first functionally grounded metrics relies on the establishment of a ground truth that will be used as the reference for measurements of **fidelity** of TCAV (the paper uses the terms "accuracy" but that doesn't apply to the definition of this paper).

The experimenters train networks with labeled data of three classes, zebra, cab and cucumber. For each corresponding training sets, there will be a literal caption of the class, that appears with a certain probability noise. For example for a zero noise probability (p = 0), the caption will consist of the word "cab", for a half probability, there is a 50% percent chance that this caption will be substituted with another random word (e.g. "carrot"). Examples of such captions can be found on Figure 1. The ground truth of the relevant concept for each concept can be approximated by using testing data with or without caption. We can thus deduce which of the caption or the image itself has been more important in classification, and evaluate the **fidelity** of the TCAV, with the quantification of the sensitivity for each of the two concepts (caption or image).

There is a human grounded evaluation based on a form of **simulatability** of the approximation of the ground truth using saliency maps. Humans are presented saliency maps and can rate the importance of image or caption on a 10 point-scale. Users performed at best random in the perception of the importance of the two concepts. The point is to demonstrate that saliency maps, which is a fairly popular explanation techniques, fails to highlight the relevance of a certain concept (image or caption) in the classification process, in contrast with TCAV.

There is a lack of application based evaluation although the paper presented an application in medical diagnosis that is useful.

Maybe a utility-performance based evaluation could have been appropriate here.

4.2 SIDU

The evaluation encompasses functional, human and application based metrics: **Fidelity**, **Robustness**, **Simulatability** and **Human Judgement**. A comparison with two other popular model-specific methods (Grad-Cam and Rise) takes an integral part of the evaluation process.

For fidelity, the correlation between model prediction and visual explanation, there are two automatic causal metrics, deletion and insertion. Respectively they apply to pixels of an important saliency region in the decision process. Deletion of pixels should force the model to change decision whereas

Technique	Evaluation		
	Functionally-Grounded	Human-Grounded	Application-Grounded
TCAV [14]	**Fidelity**.	**Simulatability**.	No evaluation.
SIDU [18]	**Fidelity** with causal tools of deletion and insertion that are measured using AUC. **Robustness**.	**Simulatability**, results are cross-examined with other XAI techniques using mathematical tools.	**Human Judgement** of medical experts.
ACE [10]	**Importance** (can be understood as a form of **fidelity**) using Smallest sufficient concepts (SSC) or Smallest destroying concepts (SDC)	**Coherency** (can be understood as a form of **comprehensibility**). **Meaningfulness** (can be understood as a form of **comprehensibility**)	No evaluation
Net2Vec [1]	**Fidelity** using IoU (Intersection over Union).	No evaluation	No evaluation
Concept Analysis with ILP [13]	**Fidelity** of concepts importance using IoU metric and first-order rules explanations using accuracy and F1 metrics	No evaluation	No evaluation

Table 3: Evaluation of 5 techniques

the model should quickly come to a correct decision with the insertion of pixels. A concrete tool for this evaluation is the **AUC** metrics for fidelity like the one described in section 3.1. Concerning deletion, the area under the curve should drop sharply as well as the predicted score.

There is functional evaluation of resilience, **robustness** against Adversarial attacks. Adversarial attacks modify source images towards different classification results without any noticeable visual change. SIDU is evaluated against a successful attack called Fast Gradient Sign Method which adds slight noise. Similar as the process for human grounded evaluation, eye tracking and XAI generated heatmaps are compared, which results will be cross-examined between different XAI techniques. The decrease of performance with the increase of adversarial noise can measure the extent of **robustness**. More simply it is also possible to measure the deviation of the saliency region with the increase of adversarial noise.

Simulatability will be based on the approximation of a ground-truth using eye-tracking of non-expert subjects.

The participants are shown random images and are asked to determine the object class that is presented. The eye fixations when the participant looks at the image for recognizing the object class will be immediately recorded. After collection, a heatmap is generated using Gaussian filter convolving, ready for comparison with the corresponding XAI generated heatmap. Finally, several statistical tools are used for cross-examinations with the other XAI techniques such as Area Under ROC Curve, Kullback-Leibler Divegence and Spearmans Correlation Coefficient. Figure 2 Shows an example of such heatmaps. This approach is effective but can be difficult and expensive to implement on a larger scale. This can be understood as a form of **simulatability** because we assess whether a human subject can perform like the explanation technique.

Application grounded evaluation of **human judgement** relies on the involvement of domain experts, ophthalmologists get on the task of retinal fundus image quality assessment. Experts will rely on their state-of-the art method to localize the exact region for prediction of the retinal fundus image quality and are presented with different samples with saliency regions with no prior knowledge of the explanation technique used. Thus, they will judge the explanation as good or bad, which will be averaged and compared with another XAI technique (RISE).

4.3 ACE

For the evaluation process, the experimenters have defined desiderata that apply specifically to a concept-based explanation:

Meaningfulness: An example of a concept, should convey a semantic meaning. Also, different individuals should recognize similar meanings to a concept. This can relate to the notion of the **Comprehensibility**, a meaningful explanation is understandable by improving the mental-model of user because he can connect computed concepts to subjectively concrete ones.

Coherency: Different examples of the same concept should be perceptually similar and should differ from examples of other concepts. We can relate it to the functional notion of stability/robustness, with distinction that it relies on human perception. Also, we can assume that the user will trust and understand better explanations with better coherency. Thus, we can consider it as a form of **comprehensibility**.

Importance: Concepts should be important. Their presence is necessary for the prediction process. This can be understood as a form of **fidelity**.

Experimenters use both quantifiable human-grounded and functionally-grounded metrics to assess the desiderata. For coherency, subjects participate in intruder detection experiments as they are presented six images, five of the same con-

cept and one intruder. The images are either hand-labeled or extracted from the ACE, for comparison. Users performed on average better with discovered concepts than with hand-labeled concepts (97% vs 99% correctness) which confirms the coherence.

To evaluate meaningfulness, subjects are presented four segments of the same concept and four random segments from images of the same class, then are asked to choose the most meaningful option. Furthermore, users are asked to describe the chosen concept with one word. Words from users are compared to verify equality or synonymy of words. An example of such a questionnaire can be seen on Figure 3.

To functionally quantify the importance of concepts, the Smallest sufficient concepts(SSC) and Smallest destroying concepts(SDC) metrics are employed. The first looks at the smallest sufficient set of concepts for correct prediction, the second at the smallest sufficient removal concepts for incorrect prediction. Measurement is straightforward as it requires to check prediction results while adding/ removing discovered concepts on a segment of standardized resolution.

4.4 Net2Vec

The evaluation first relies on better performance of **fidelity** for discovering concepts and then on a demonstration of the ability to build a powerful mean of **comprehensibility** based on an understanding of the relations between concepts.

To measure the effectiveness of a filter or group of filters to produce a segmentation a metrics called **IoU**(Intersection over Union) is used. It computes the intersection over union difference between segmentation masks produced by the filter and the ground-truth segmentation masks.

Using this metric, it has been found that filters can reach a four-fold improvement of IoU of single filters for simple concepts such as color. It is also found that the number of filters can result in saturation and sub-optimal results.

Finally, the paper demonstrates that, by learning the weights of different filters, it is possible to derive a conceptual embeddings from visual data. Arithmetic can be built upon and a distance formula can be derived to build a network space of concepts and think of them in terms of similarity. It is also possible to make vector operations and sum or subtract concepts from one another and discover which concept is closest to this concept. For example, “tree” minus “wood” is closest to the concept of “plant”. All of this can reinforce the desideratum of **comprehensibility** or **trust** but no rigorous evaluation using, for example human subjects, has been conducted.

4.5 Combining Concept Analysis with ILP

The evaluation process is quantified and functionally based, it is based on the **fidelity** of the concept models and of the symbolic explanations in relation with a ground truth.

Evaluation is conducted with three different architectures and a generated dataset called as Picasso Dataset. This dataset consists of human faces where the nose, eyes and mouth are either rearranged or in the correct position, while setting the rest of the face as a homogeneous skin tone. The positive class consists of faces with correct disposition facial features, the negative one of incorrectly disposed ones. Figure 4 Gives a view of such samples.

For evaluation, the experimenters determine the best ensembled detection concept vectors for MOUTH, EYES and NOSE concepts among layers. Training was made with segmentation masks. The goal is to predict whether the kernel window intersects with an instance of the concept at each activation map pixel, with a certain threshold. This is the intersection encoding, implemented with convolution. Then, the IoU metric is used, between intersection encoded masks and detection masks, to quantify the relevance of an ensemble, in addition with a cosine distance.

As for the logical explanations, some samples of the Picasso data-set were used to train the ILP model (Aleph). First, they use the masks of the samples drawn in the previous step to extract the facial features information. They build the background knowledge and extract the spatial features between the found parts. Aleph can thus induce a theory of logic rules for a trained network. The fidelity of explanations is measured in relation with the initial black-box model, which binary outputs serve as ground-truth. The explanation rules for each example are used as binary classification model. The faithfulness of the later results is measured with Accuracy and F1 metrics.

5 Discussion

There is a lack of standards regarding the evaluation process. Terms can be used interchangeably to refer to the same thing, e.g. accuracy and correctness, faithfulness and fidelity. This is further confirmed by the evaluation of the ACE technique that relies on the definition of new desiderata that can be understood in terms of faithfulness or simulatability.

Also, the majority of the examined evaluation of model-specific techniques don’t seem to follow a rigorous and exhaustive procedure, several criteria have not been assessed. Additionally, SIDU is the only technique where three different forms of metrics are evaluated: functionally, human and application grounded. SIDU is also the only paper that directly evaluates the robustness of the technique, in the context of an adversarial attack. This is lacking in the four other papers, an evaluation of the resilience of techniques towards adversarial attacks is one of the main neglected aspect of evaluation, in particular in the form of the robustness and stability criteria. There seems to be a lack of a more holistic approach to the evaluation, papers are usually focused on one to three main criteria.

Fidelity is the only criterium to be evaluated by the five papers. The combined concept analysis and ILP evaluation only focuses on faithfulness. Making sure that an explanation is reflective of reality seems to be the main goal of explanation techniques. The nature of the explanations that are based on concept evaluation, visualisation and simple first-order logic explanations are evocative and can make clear mental-models. Thus, quality of explanation, comprehensibility seem to be taken for granted.

In general the examined papers present a novel technique for a certain type of explanation, in this case concept-based, and how that explanation performs better in this particular

context, usually in comparison with other similar techniques. The Net2Vec makes it clear by being more of a demonstration of a technical ability than a quality of explanation that directly relates to a user.

Finally, material obstacles to evaluation were discovered. The SIDU paper [18] explains that conducting human and application based evaluation is costly, it is therefore not always materially possible to conduct a thorough evaluation of a technique. The description of a case application of the technique is not systematically mentioned with the exception of TCAV and SIDU. This would an important aspect of the evaluation, as it gives context and reasons for the particular conducted assessment and gives direction to the readers looking for an explanation technique.

6 Future Work

Future work can be evaluations specialized in the application and human-grounded evaluation of XAI on a variety of techniques. This could be done in a bigger environment of experimentation tat provides the means to face the costs of evaluation requiring human subjects. So far, this cost has prevented a proper evaluation of the user perceivable quality of explanations.

Promising further research, might also focus on the evaluation of a specific metric in the evaluation of the compared techniques. For example, experimenters who possess a rigorous robustness evaluation method could provide a more comprehensive view of the techniques abilities.

Moreover, the new desiderata defined for the SIDU evaluation, can be explored and their importance investigated. Then, they could be applied to the evaluation of other techniques.

Also, the introduction of standards of evaluation as well as proposals for protocols of evaluation can be promising. For example, a paper could argue for a standard that requires both fidelity and robustness to be evaluated while suggesting some evaluation tools.

7 Conclusion

In this paper the different available tools and criteria for the evaluation of XAI techniques have been investigated. Then, the evaluation process of five state-of-the-art model specific techniques is surveyed. It has been discovered that there is usually a lack of rigorous and holistic assessment of metrics in the evaluation, which can make fragile certification of the explanation techniques. This is reinforced by the material constraints that might face evaluation endeavours. Furthermore, a lack of standards can also be a factor, it also explains the mixing between certain evaluation tools terms and their meanings for explanation.

8 Responsible Research

The research conducted lays its foundation on a literature review, that takes sources from recent scientific journals and conferences about XAI, mainly from IEEE and arxiv. All the information, summaries and implications that have been derived are therefore reproducible and verifiable as one can look at the citations. I have made my best effort to be honest and faithful about the source material, and analyze diverse perspectives on the topic. However, there is still a subjective part due to the diverse nature of surveys on evaluation methods as well as a substantial part of the work that focuses on a comparison between the evaluation of five different techniques, and a discussion about some neglected aspects. I have therefore made the best effort to be the closest to a scientific consensus regarding the summary of the evaluation tools, adopting a taxonomy for the different evaluation tools that is the most prevalent in the literature. Furthermore, I have tried to be the most objective about the judgement of the evaluation process of the five techniques, ensuring that my perspective is grounded on a widely accepted framework of the evaluation of XAI.

References

[1] Ghorbani A., Wexler J., Zou J.Y., and Kim B. Towards automatic concept-based explanations. volume 32, pages 9273,9282, 2019. URL: papers.nips.cc/paper/9126-towards-automatic-concept-based-explanations.

[2] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). volume 6, pages 52138–52160. IEEE, 2018. doi:10.1109/ACCESS.2018.2870052.

[3] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. 2018. doi:arXiv:1806.08049.

[4] Fang Chen, Amir H. Gandomi, Jianlong Zhou, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 2021. doi:doi.org/10.3390/electronics10050593.

[5] Saikat Das, Namita Agarwal, Deepak Venugopal, Frederick T, Sheldon, and Sajjan Shiva. Taxonomy and survey of interpretable machine learning method. 2021. doi:10.1109/SSCI47803.2020.9308404.

[6] Brittany Davis, Maria Glenski, William Sealy, and Dustin Arendt. Measure utility, gain trust: Practical advice for xai researchers. IEEE, 2020. doi:10.1109/TREX51495.2020.00005.

[7] T. De, A. Mevawala, and R. Neman. An explainable ai powered early warning system to address patient readmission risk. 2021. doi:10.1109/NBEC53282.2021.9618766.

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017. doi:arXiv:1702.08608v2.

[9] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. In *Communications of the ACM*, volume 63, page 68–77, 2020. URL: doi.org/10.1145/3359786, doi:10.1145/3359786.

[10] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by lters in deep neural networks. pages 8730,8738, 2018. URL: doi.org/10.1109/CVPR.2018.00910.

[11] Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha. An empirical evaluation of ai deep explainable tools. IEEE, 2020. doi:10.1109/GCWkshps50303.2020.9367541.

[12] Ambreen Hanif, Xuyun Zhang, and Steven Wood. A survey on explainable artificial intelligence techniques and challenges. In *IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, 2021. doi:10.1109/EDOCW52865.2021.00036.

[13] Rabold J., Schwalbe G., and Schmid U. Expressive explanations of dnns by combining concept analysis with ilp. pages 148–162, 2020. URL: doi.org/10.1007/978-3-030-58285-2.

[14] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). volume 80, pages 2668,2677, July 2018. doi:arXiv:1711.11279.

[15] Xiao-Hui Li, Yuhan Shi, Haoyang Li, Wei Bai, Caleb Chen Cao, and Lei Chen. Quantitative evaluations on saliency methods: An experimental study. 2020. URL: doi.org/10.1145/3447548.3467148, doi:arXiv:2012.15616v1.

[16] Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. Evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. 2020. doi:arXiv:2009.10639v1.

[17] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. In *ACM Transactions on Interactive Intelligent Systems*, volume 11, pages 1–45, 2021. URL: doi.org/10.1145/3387166, doi:arXiv:1811.11839.

[18] Satya Muddamsetty, Mohammad Jahromi, Andreea Ciontos, Laura Fenoy, and Thomas Moeslund. Introducing and assessing the explainable ai (xai) method: Sidu. January 2021. URL: arxiv.org/abs/2101.10710, doi:arXiv:2101.10710.

[19] A. Rawal, J. Mccoy, D. B. Rawat, and R. Amant B. Sadler. Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives. In *IEEE Transactions on Artificial Intelligence*. doi:10.1109/TAI.2021.3133846.

[20] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concept. 2021. doi:arXiv:2105.07190v2.

[21] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. In *IEEE Transactions on Neural Networks and Learning Systems*, volume 32, pages 4793–4813. IEEE, 2021. doi:10.1109/TNNLS.2020.3027314.

[22] Orcun Yalcin, Xiuyi Fan, and Siyuan Liu. Evaluating the correctness of explainable ai algorithms for classification. 2021. doi:arXiv:2105.09740v1.

9 Appendix

Figure 1: Examples of captions, for a ”cab” and ”cucumber” concepts

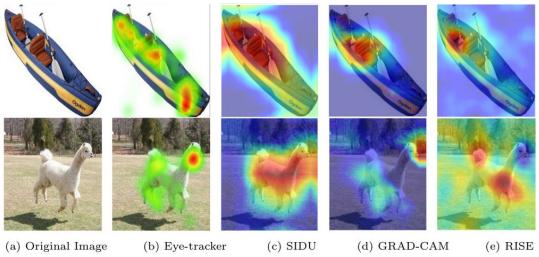

Figure 2: Heat maps resulting from eye-tracking and different vizualisation explanation techniques

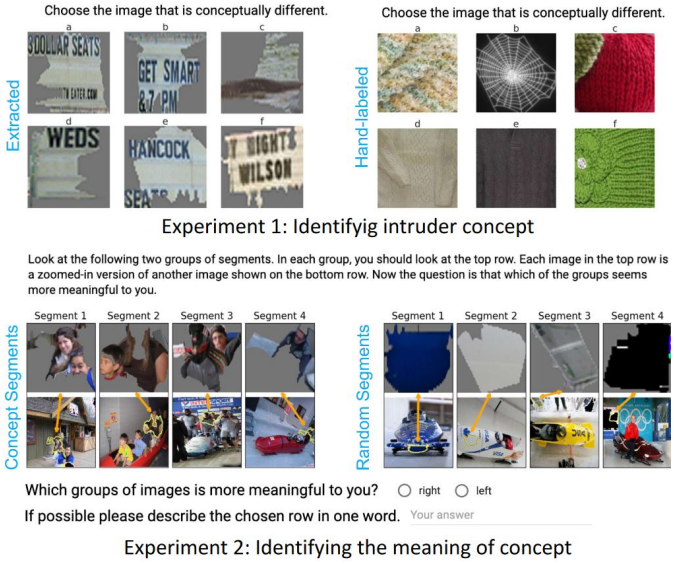

Figure 3: Example of questionnaire to measure meaningfulness

Figure 4: Left are positive class examples, facial features are disposed normally. Right is negative