Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 1, 133-142 (2024)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

A Novel Metric for Evaluating the Stability of XAI Explanations

Falko Gawantka[*,1], Franz Just[1], Marina Savelyeva[1], Markus Wappler[2], Jörg Lässig[1,2]

[1] *University of Applied Sciences Zittau/Görlitz, Faculty of Electrical Engineering and Computer Science, Görlitz, 02826, Germany*

[2] *Fraunhofer IOSB, Advanced System Technology (AST), Görlitz, 02826, Germany*

ARTICLE INFO

Article history:
Received: 15 November, 2023
Revised: 16 January, 2024
Accepted: 17 January, 2024
Online: 21 February, 2024

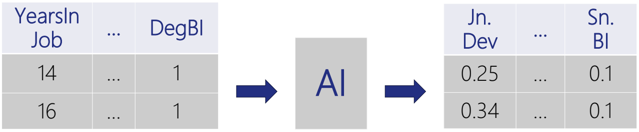
Keywords:
eXplainable AI
Evaluation
Stability of explanations

ABSTRACT

Automated systems are increasingly exerting influence on our lives, evident in scenarios like AI-driven candidate screening for jobs or loan applications. These scenarios often rely on eXplainable Artificial Intelligence (XAI) algorithms to meet legal requirements and provide understandable insights into critical processes. However, a significant challenge arises when some XAI methods lack determinism, resulting in the generation of different explanations for identical inputs *(i.e., the same data instances and prediction model). The question of explanation stability becomes paramount in such cases. In this study, we introduce two intuitive methods for assessing the stability of XAI algorithms. A taxonomy was developed to categorize the evaluation criteria and the ideas were expanded to create an objective metric to classify the XAI algorithms based on their explanation stability.*

1 Introduction

As artificial intelligence (AI) becomes increasingly integrated into people's lives, a notable trend can be observed: automation of processes and domains that significantly impact the well-being and future prospects of individuals. For example, the initial screening of job applicants can be considered. Given the substantial human resource expenditure involved in this process, a compelling case can be made for its automation. Candidate scores can be calculated using artificial intelligence, as demonstrated in previous research [1]. Refer to Figure 1 for more details. Nevertheless, the decision-making process utilized by the underlying machine learning model presents a challenge. The accumulation of applicant information often results in a lack of transparency, a concern shared by HR managers, companies, applicants, and regulators alike.



YearsIn Job	...	DegBI
14	...	1
16	...	1

Jn. Dev	...	Sn. BI
0.25	...	0.1
0.34	...	0.1

Figure 1: The job application pipeline is depicted here. On the left, the input values predicted by the ML model are displayed, while on the right, the corresponding results are presented. Transparency issues arise when the information density is reduced to a single numeric value. This lack of transparency makes it neither intuitive nor comprehensible for the stakeholders involved. The question of: „What impact does feature x had on the prediction?" arise.

One viable approach to address this concern is the integration of explanation algorithms, often referred to as XAI algorithms. This strategy enables visualization of model decisions in an understandable way. Additionally, both the transparency and the validity of the predictive model can be improved, ensuring that the needs and expectations of the relevant stakeholders are met.

However, it is essential to recognize that not all explanatory algorithms can be used without restrictions in high-stake areas [1]. This journal paper extends the work originally presented in IEEE ICCI*CC'22 [1]. The stability of the explanations was investigated through different XAI approaches and ranked according to their results. This was achieved by quantifying the feature importance values (FIVs). In this work, the state of the art for the evaluation of XAI explanations is presented, and the taxonomies found are expanded or modified in terms of the aspect: *stability*. The corresponding research questions are as follows.

- RQ1: What does the *stability* of an explanation mean?
- RQ2: How can the explanation of an XAI algorithm be measured in terms of *stability*?

By that, the following hypothesis can be stated: „High *stability* of an explanation is observed when the explanation undergoes minimal changes in response to minor variations in unimportant features of the data instance."

[*] Corresponding Author:University of Applied Sciences Zittau/Görlitz, Faculty of Electrical Engineering and Computer Science, Görlitz, 02826, Germany, Email: falko.gawantka@hszg.de

[1] This domains could harm people or the future of a person, for instance here the HR use case represents an automated decision that has an impact on a person.

https://dx.doi.org/10.25046/aj090113

For example, in the case of an HR applicant selection process, a change in the input features (i.e. 14 vs 16 *YearsInJob*) should not produce a substantially different explanation.

The paper is organized as follows. In section 2, metrics and evaluation techniques for XAI algorithms are presented, and the foundations of XAI and representative methods are discussed in detail. In section 3, an approach to quantify explanation stability is introduced. The details and limitations of previous approaches serve as a basis for a new scoring metric. The results are shown in section 4. Potential future research directions are outlined in the discussion in section 5.

2 Related Work and Background

For a deeper understanding and examination of XAI algorithms, an overview of evaluation taxonomies, methodologies, and the underlying calculation models is provided, forming the theoretical foundations. In addition to that, some background information are provided to enhance the understanding of XAI. This work aims to be considered as best practices in the selection process of XAI algorithms for real-world use cases, with the objective of uncovering a comprehensive understanding of the model with high *stability* to the respective explanation of XAI.

2.1 Related Work on XAI evaluation

To explore the stability of an explanation result, this subsection first defines the relevant properties/criteria. It then provides an overview of the measuring techniques and presents the mathematical foundations.

2.1.1 Notions by different Taxonomies

In the work of [2], a comprehensive taxonomy is introduced, a distinction is made between *the user aspect*, *the explanation aspect* and *the model aspect*. The focus of this work lies on *the explanation aspect*. In this aspect there are two further sub-criteria defined by [2], which have to do with stability: *Identity* and *Separability*. *Identity* is defined by: “ identical instances should have identical explanations” and *Separability* is defined by: “non-identical instances should not have identical explanations” [2]. To describe these two aspects, one can think of the ends of a scale.

A paper cited by [2] was that of [3] that defines *Stability* as: It represents how similar are the explanations for similar instances. While consistency compares explanations between different models, stability compares explanations between similar instances for a fixed model. High stability means that slight variations in the feature values of an instance do not substantially change the explanation, unless these slight variations also strongly change the prediction.

Nonetheless, a lack of stability can also be created by non-deterministic components of the explanation method, such as a data sampling step. Regardless of that, high stability is always desirable. This definition brings another dimension into play. The author in [2] speaks only of “identical instances”, and [3] also mentioned the predictor by constructing a context around the instance and saying that there is also a predictor and an instance.

Following the definitions in [2], the author in [4] describes in the work the so-called *Co-12 properties*. The *Consistency* property coincides with the *Identity* criterion from [2]. *Consistency* is defined as: “Identical inputs should have identical explanations” [4]. The term inputs can be defined by the data instance as well as the ML model. In this context, the model’s perspective is integrated, with the rest remaining consistent with the definition in [2]. And there is also a second aspect in [4], which can be equated with *Stability* in [3]. The term *Continuity* introduced by [4] is defined as follows: “Similar inputs should have similar explanations” [4]. The use of similar inputs could mean both the model view and the data view.

Besides the *Stability* notion by [3] and the term *Continuity* by [4] there is a another work [5], where the author measured the *Stability* of XAI algorithms. The underlying assumption was that: If the inputs are almost identical, then it is expected that only minor changes appear in the explanation. This can be used as a scoring metric for the stability of an XAI algorithm. The term *almost identical* refers hereby to *similar* and therefore matches the understanding of *Stability* by [3] and the notion *Continuity* from the work of [4]. The author in [5] gives no explicit definition, but: “similar instances and an identical model should have similar explanations”, could be suggested as a common understanding for the property of *Stability*. Another property or criterion investigated in [1] was the *Reproducibility/Stability of the explanation*. This refers to an aspect in which an identical model predicts an identical data instance several times (i.e., 10 times, 100 times, and 1000 times). In the context of the reproducibility of explanation stability, *Reliability* would be a more appropriate term than the one chosen by the author.

The entire notions of the taxonomies are grouped and summarized on the *Stability Scale* and could be seen in Figure 2.

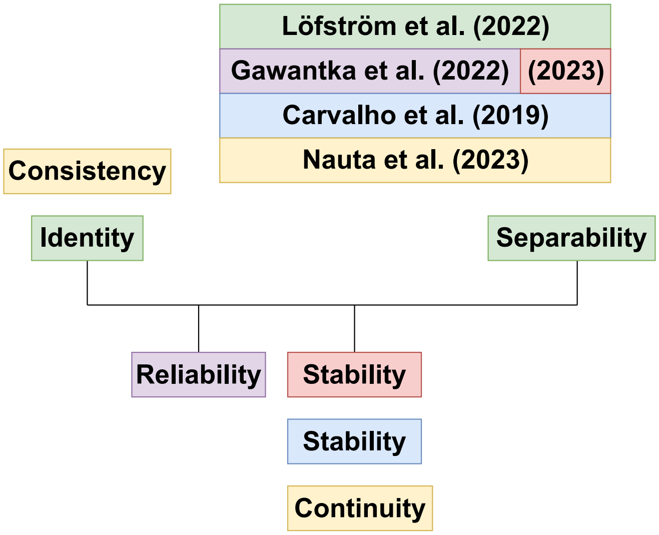


Figure 2: Stability Scale: The scale visualize the different terms and group them on the scale to show what terms could be used interchangeable.

The focus of this work lies in unifying different schools and terms, proposing concrete measuring techniques, and establishing a strong foundation from the mathematical perspective of XAI evaluation metrics.

2.1.2 *Quantifying XAI through Feature Importance Values*

A formal introduction into machine learning is initiated, drawing inspiration from a more detailed section in [6]. A comprehensive introduction to artificial intelligence is available in the textbook [7]. The notion of explanation method is taken and adapted from [8].

Feature importance values are used to quantify the explanations. These are numerical scores that help to evaluate the relative significance of individual features of the model's predictions, thus offering transparency and interpretability. It is important to note that the choice of the feature importance method matters and that different methods may lead to varying results. Therefore, careful consideration is needed to select the most appropriate method for a particular application. Adopting the notions presented in subsubsection 2.1.1 to the feature importance values will provide evaluation metrics to support the selection of data-driven methods.

In the context of supervised learning, XAI and the feature importance values will be considered. This is based on two fundamental components: the model and the data. Given an input domain X and an output domain Y, supervised learning aims to predict an output $y \in Y$ for given $x \in X$. In the example of classifying job applicants, the input domain X reflects the data structure of the applicant profile, while Y encodes the job profiles. The prediction mechanism is described through a mathematical model that can be deterministic or probabilistic.

A *deterministic model* is a function $f : X \rightarrow Y$ that predicts exactly for a given input $x \in X$ an output $f(x) \in Y$. This can be a classification prediction, in case Y is a discrete set, or this can be a regression prediction, if Y is a continuous set. For example f assigns a job profile $y \in Y$ to each applicant for a job $x \in X$.

A *probabilistic model* is a probability distribution $p : X \times Y \rightarrow [0, 1]$. It can be recognized as a full distribution, giving the probability of assignment of $x \in X$ and $y \in Y$ by $p(x, y)$, or it can be recognized as a conditional distribution, providing the probability of output $y \in Y$ for chosen $x \in X$ by $p(y|x)$. In the latter case for example, p would encode the probability that an applicant's profile $x \in X$ will be classified as job profile $y \in Y$.

In either deterministic model or probabilistic model, the underlying function is usually parameterized. Training the model means optimizing the parameters to achieve good predictions, based on a data set $\{(x^1, y^1), \dots, (x^m, y^n)\} \subseteq X \times Y$.

The input data are commonly structured data $X = X_1 \times \cdots \times X_n$ with the notion of a *feature* being applied context-dependent to any primitive domain X_k or to any data entry $x_l \in X_k$ or instance $x \in X$. The output $y \in Y$ is called the label.

In the remainder of the paper, all models are considered to be deterministic. The set of all (deterministic) models is denoted by $\mathcal{F} = X \rightarrow Y$.

Having a (trained) model f, an explanation method $\Phi : \mathcal{F} \times (X = (X_1, \dots, X_n)) \rightarrow \mathbb{R}^n$ attributes an importance score $\Phi(f, x)_i$ to each feature x_i describing its impact on prediction $f(x)$. The range of feature importance values is method-dependent. Common choices are $(-\infty, \infty)$ for SHAP, $(0, 1)$ for Gini importance, $(0, \infty)$ for Gain, $(-1, 1)$ for Correlation Coefficients.

Using this notion, *Consistency* means that $\Phi(f, x)$ evaluates to the same value, regardless of how often it is calculated. Obviously, this is always fulfilled, making the term irrelevant in the context of feature importance values. But still, a similar metric might be useful, as certain XAI methods, e.g. surrogate explainer methods, train a model and try to approximate the original ML model, e.g. LIME is such a method. In fact, each LIME instance $\Phi(f, x)$ might behave differently depending on the training data points used. Demanding that explanations at least do not deviate much would be a natural extension of *Consistency*.

Thus, in [1], another evaluation criterion known as *Reliability* is mentioned here. For every call of Φ an element is added to the result set $\{\Phi^{(r)}(f, \cdot)\}_r$. One can expect that a statistical quantity that measures the deviation, such as the length of confidence intervals, will decrease with increasing r. However, its rate of decrease provides a measure for *Reliability*.

The definition of *Reliability* is shown in the following equation:

$$\{\Phi^{(r)}(f, \cdot)\}_r = \left\{ \begin{pmatrix} fiv_{i=1} \\ \vdots \\ fiv_{i=s} \end{pmatrix}_{r=1}, \cdots, \begin{pmatrix} fiv_{i=1} \\ \vdots \\ fiv_{i=s} \end{pmatrix}_{r=c} \right\} \tag{1}$$

In Equation 1 the result of a single call from Φ is shown as vector. For every call r an explanation vector ($\vec{F}$) is calculated, so $fiv_i \in \vec{F}$. The result is a set of vectors that represents a matrix.

Based on this aspect, the mean over the 1st feature importance value from r-repetitions could be defined by:

$$\mu_{F_{(i=1)}} = \frac{1}{c} \sum_{r=1}^{c} fiv_{(i=1)r} \tag{2}$$

Therefore the variance could be calculated by the mean:

$$\sigma^2_{F_{(i=1)}} = \frac{1}{c} \sum_{r=1}^{c} (fiv_{(i=1)r} - \mu_{F_{(i=1)}})^2 \tag{3}$$

The standard deviation for a single feature importance value is then calculated from:

$$\sigma_{F_{(i=1)}} = \sqrt{\sigma^2_{F_{(i=1)}}} \tag{4}$$

Finally, the deviation can establish the foundation for the confidence interval of a single feature importance value.

$$CI_{F_{(i=1)}} = \mu_{F_{(i=1)}} \pm (z = 1.96) \cdot \frac{\sigma_{F_{(i=1)}}}{\sqrt{c}} \tag{5}$$

In the previous equations, an approach was presented to calculate the confidence interval of the single feature importance value of the feature importance vector $\vec{F}$. When this is done over the entire feature importance vector and then the mean is built on top of it, this matches the metric from [1]. This calculation process is denoted by $R(x)$.

The intuitive definition of *Stability* involves small perturbations of an instance that result in only minor changes to the explanations under the same model. To be numerically tangible this definition has to be measurable, hence the notion of correlated small changes needs to be quantified. In [9] the explanation methods SHAP and LIME are evaluated in terms of Lipschitz continuity, using the best Lipschitz constant as a stability measure. This approach is generalized while keeping it local (depending on the chosen data point). Furthermore, it will be demonstrated that the different stability measures previously used fall under this umbrella.

Lipschitz continuity can be defined in a general sense for metric spaces as follows: A function $f : X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is called *Lipschitz continuous* if there is a real constant $L \geq 0$ (referred to as *Lipschitz constant*) such that for all $x, \tilde{x} \in X$,

$$d_Y(f(x), f(\tilde{x})) \leq L d_X(x, \tilde{x})\,. \tag{6}$$

This is, the distance of the images of f is bounded by the distance of the respective arguments up to a constant factor. In case of local explanation methods, e.g. as SHAP and LIME are, an instance-independent constant is not attainable. Therefore, we will base the definition on a data point-dependent Lipschitz factor within a neighborhood of the data point.

Let X be an input domain, equipped with the metric d_X. Let Y be an output domain equipped with the metric d_Y. Let $f \in \mathcal{F} = X \to Y$ be a model, and $\Phi : \mathcal{F} \times X \to X$ be an explanation method. An explanation $\Phi(f, x)$ can be termed *ε-stable* if there is a local real constant $L(x) \geq 0$ such that

$$d_Y(\Phi(f, x), \Phi(f, \tilde{x})) \leq L(x) d_X(x, \tilde{x}) \tag{7}$$

for all $\tilde{x} \in N_\varepsilon(x) = \{\tilde{x} \in X : d_X(x, \tilde{x}) \leq \varepsilon\}$. The Lipschitz factor $L(x)$ can be seen as a stability measure. The lower $L(x)$, the smaller the deviation, the higher the stability. Finding small Lipschitz factors can be done using the Monte Carlo method. Examples from the literature will be provided that specify the notion of stability.

Example 2.1. In [9] the metrics are induced by norms, resulting in $N_\varepsilon(x) = \{\tilde{x} \in X : \|x - \tilde{x}\| \leq \varepsilon\}$. The best Lipschitz constant is given by

$$L(x) = \max_{\tilde{x} \in N_\varepsilon(x) \setminus \{x\}} \frac{\|\Phi(f, x) - \Phi(f, \tilde{x})\|}{\|x - \tilde{x}\|}\,. \tag{8}$$

An approximation can be done by sampling with $\{x_1, \ldots, x_n\} \subset N_\varepsilon(x)$.

Example 2.2. In [10] an evaluation metric for XAI methods called Max-Sensitivity was introduced. The concept is taken up by [8]. To derive the term, the inequality is revisited (7) and relaxed to

$$d_Y(\Phi(f, x), \Phi(f, \tilde{x})) \leq \varepsilon L(x) = L'(x)\,. \tag{9}$$

Now, again using norms as distance metrics, $L'(x)$ can be obtained by

$$L'(x) = \max_{\tilde{x} : \|x - \tilde{x}\|_\infty \leq \varepsilon} \|\Phi(f, x) - \Phi(f, \tilde{x})\|_2\,. \tag{10}$$

This is in fact the Max-Sensitivity metric. Again, it can be approximated by sampling.

Example 2.3. The work [5] considers a similarity measure by first generating data points by perturbing a feature around 1% and then evaluating the ratio of original feature importance value and perturbed feature importance value for that feature. Furthermore, the feature importance values are normalized so that they form a probability distribution.

Modifying this approach by allowing perturbations of at least ε, which bounds the values around one percent, allows to extend this measure to

$$S_i(x) = \max_{\tilde{x} \in N_\varepsilon(x)} \left| 1 - \frac{\Phi(f, x)_i}{\Phi(f, \tilde{x})_i} \right|. \tag{11}$$

Therefore, $S_i(x)$ is the best similarity to the feature i that all $\tilde{x} \in N_\varepsilon(x)$ fulfill. As a combined score for all features could be

$$S(x) = \max_{\tilde{x} \in N_\varepsilon(x)} \frac{\|\Phi(f, x) - \Phi(f, \tilde{x})\|_\infty}{\|\Phi(f, \tilde{x})\|_\infty}\,. \tag{12}$$

On one hand, this is similar to Lipschitz-based stability criteria; on the other hand, it effectively demonstrates that the criteria proposed in [5] normalize the feature importance values.

2.2 Background of XAI Taxonomies

At this point, eXplainable Artificial Intelligence is introduced as a field of study within the domain of artificial intelligence, along with an explanation of its associated focuses and terminology.

In [11, p. 3511] the author defines XAI as „the study of explainability and transparency for socio-technical systems, including AI.“ The 2019 XAI Taxonomy by [12] introduces important terms in explainable artificial intelligence. Later, in [13] used this taxonomy as a decision-making tool to choose the appropriate explanatory algorithms for the IBM AIX 360 tool. The tree structure of the taxonomy of [13] was transformed into a tabular format and can be seen in Table 1. The focus of this work is to provide an overview of explanatory models, which is highlighted in the table. The *data* leaf is not shown in the table and the interactive path is actually empty. The explainability of the *model* has a *local* and a *global path*; that is, there is a distinction between explaining the data sample and explaining the predictive model. Local interpretations of a single data instance can be done *ante-hoc*, i.e., the predictor is so comprehensible that a closer look serves as an explanation, for example, when the ML model is a simple decision tree. The other path, called XAI *post-hoc* methods, needs at least one execution of the predictor to produce information about the decision process of the ML model. The *global* model explanation is divided into two parts. The *direct* methods have the notion of direct interpretable and could be seen as *ante-hoc*. *Post-hoc* refers, in contrast to that, to algorithms that can create an explanation.

Due to rapid development in the field of XAI, the results of Table 1 are supplemented by other essential terms. The simple XAI taxonomy of [14], as well as the more sophisticated categories of [15], served as the basis for the following taxonomy, which has three sub-categories. The first sub-category is shown in Table 2.

Table 1: This IBM XAI Taxonomy is derived from the decision tree presented in [13], and serves as a guide to find a suitable XAI method. This overview focuses specifically on the model explanation aspect and does not present specific algorithms.

model					
local			***global***		
ante-hoc	***post-hoc***		direct	***post-hoc***	
	samples	features		surrogate	visualize

In Table 2, the distinction between XAI algorithms is made based on the creation of the explanation. For example, *Local Perturbation* methods try to modify the input and find important features as well as their corresponding feature importance values [14, 15]. The methods in *Leveraging Structure* use internal information from the ML model, such as gradients, like *Backprop* in the paper by

[14]. *Meta Explanations* abbreviated with *Meta-Expl.* exploit explanations from different explanatory methods, and *Architecture Modification* abbreviated with *Arch. Mod.* algorithms alter the predictor with the goal of finding simpler representations [15]. *Examples* Methods derive explanations from input data and use the output as an interpretation for similar inputs.

Table 2: This table presents the functional part by the XAI Taxonomy according to [14] and [15]. The methods are classified based on their functional interaction with the predictor model.

funtional approach				
Local Perturbation (Perturbation)	Leveraging Structure	Meta-Expl.	Arch. Mod. (Backprob)	Examples

Another classification uses the output of the explanation approaches, which is shown in Table 3. The first sub-category is the weighting of feature importance values that provides information on which features have an influence and how great this influence is. *Surrogate Models* try to approximate more complex models by focusing only locally or by generally using *ante hoc* explanatory models. In addition, example-based methods are also used here. [15]

Table 3: The scope of this part from the introduced XAI Taxonomy according to [14] and [15] is focused on the results that different explanation models provide. As it can be seen there are *Feature Importance* as well as *Example* explanations and *Surrogate Models* that can explain black box models.

result approach		
Feature Importance	Surrogate Models	Examples

The conceptual approach, shown in Table 4, is the last category introduced by [15] and covers all aspects that were introduced by [14] in 2020. Further breakdowns are made into the *usage*, the *scope* of the explanation, and the *other dimensions*. The *usage* includes procedures that can be explained from the outside or XAI algorithms that provide a *model agnostic* (*model-agno.*) or *model specific* (*model-spec.*) explanation after the predictor model. The *scope* describes the range of what is explained, e.g., the complete/*global* ML model or individual data instances. *Other dimensions* include outputs, among others [15]. IBM researchers suggest the following approaches for a feature-based explanation: CEM [16], LIME [17], and SHAP [18]. To understand the decision underlying the model, it is not enough to explain the sample (data instance), but it is fundamental to explain the impact of each feature on the decision. Based on the explanations for each feature, it may also be possible to identify bias. Considering the subject of *evaluation*, it is necessary to generate nearly equal outputs, falling into the sub-category of *Feature Importances*. In the proposed use case shown in section 1, and given the axiom that explanations are an advantage of decision support systems (DSS), the easiest way to implement this is to use *local post-hoc* Methods. A summary of suitable methods and their limitations is presented.

Table 4: This view of the XAI taxonomy from [14] and [15] highlight the explanatory capability of XAI methods. The main questions addressed here are: what is being explained and whether the explanation can be obtained from the XAI method before the machine learning model makes predictions. Additional information related to the XAI approach and the problem domain is also provided.

conceptual approach						
stage (usage)			***scope***		***other dim.***	
ante-hoc	***post-hoc***		glo.	loc.	res.	issues
	model-agno.	model-spec./ (intrinsic)				

2.3 Background of XAI Algorithms

In this sub-section are reviewed three XAI methods such as LIME (Local Interpretable Model-Agnostic Explanations by [17]), SHAP (SHapley Additive exPlanations by [18]) and CIU (Contextual Importance and Utility by [19]). To describe these algorithms, two definitions are needed, namely deterministic and nondeterministic algorithm.

An algorithm whose behavior depends entirely on the input data is called *deterministic*. Thus, processing the same input data always leads to the same result. A *nondeterministic* algorithm is an algorithm that specifies several ways to process the same input data - without any specification of which option will be chosen, which can lead to either the same or different output [20].

2.3.1 LIME

The idea behind LIME is to consider the local model as a black box model. The mode of operation of LIME is based on perturbing an original data point as input into the black box model and using the resulting predictions to train an interpretable surrogate model, which locally approximates the predictions of the black box model. The explanation provided by LIME is defined by:

$$\xi(x) = \underset{g\in G}{\text{argmin}}\ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{13}$$

In Equation 13, ξ is the explanation of instance x, which is obtained through an optimization task. The function g is an interpretable local model, and G is a class of potentially interpretable models. The function f is the original predictor and π_x defines the radius of the neighborhood around instance x. $\mathcal{L}$ is the loss function that measures the accuracy of the prediction of the instance x with respect to the interpretable model g and the original prediction of f in the area of π_x around the original prediction. Ω is a complexity measure of g and serves as a penalty function.

LIME calculates feature importance values that show the contribution of each feature for and against a prediction in a certain class. LIME values are numerical, where a negative numerical value indicates that this feature is not in favor of the prediction. On the other hand, a positive numerical value shows that this feature has a positive influence on the prediction.

The advantages of LIME are the simple explanations diagrams and the ability to process various types of input [21]. However, LIME has some disadvantages: it provides only local explanations and does not show this to the end user. The interpretation of the importance plots is often hard to understand by nonexperts. The algorithm has consistency issues and can be attacked by adversarial examples [21].

The consistency issues with LIME are not only related to minimal changes in features but also to the fact that such changes can lead to completely different explanations. In previous experiments with LIME, it has been shown that by iterating with LIME over the same data instance 10, 100 and 1000 times, the inconsistency is also present [1]. These results and [22] show that the LIME algorithm is nondeterministic.

2.3.2 SHAP

The idea of the SHAP algorithm has its origins in the game theory. Calculate the extent to which a coalition (set of features) contributes or does not contribute to a particular classification based on the so-called Shapley values. The implementation used was the implementation by [18], known as *SHap Additive exPlanations*. The following definition describes the generation of explanations by the algorithm [18, 23]:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{14}$$

The function g describes the explanatory model and z' describes the data instance to be interpreted. The variable, z' may have only a subset of all features. The explanation is generated by a linear model, where $\phi_i \in \mathbb{R}$ and z'_i are either zero or one, to represent the presence or absence of a value from the feature set z' [18].

The author in [18] presents various explainer models in the framework, and the model used in [1] was the kernel explainer. The computational model approximates Shapley values with perturbations of z'. Thus, the complexity problem of computing Shapley values could be solved. SHAP also generates feature importance values such as LIME. A higher FIV of SHAP means that it contributes more to a prediction than a smaller one. A negative feature value argues against a prediction in a class and a positive value argues for a particular prediction.

SHAP also has some advantages: first, there is the sophisticated mathematical model for greater consistency and accuracy [17, 21, 23]. Besides that, the better intuition about feature weights, which is more related to humans, is another reason mentioned in [21]. Even though SHAP comes in hand with several advantages, the algorithm is non-deterministic [24], the approximation time of the Kernel SHAP is an issue [21, 23] and SHAP is also vulnerable to adversarial attacks [21].

2.3.3 CIU

The third model-agnostic algorithm is based on Decision Theory, more specifically, based on the subdomain of Multiple Criteria Decision Making (MCDM). In contrast to LIME and SHAP, this approach distinguishes between the measured importance and the utility of an attribute. Based on the relevance of the features, the focus lies on contextual importance (CI). This is described in [19, 25] as follows:

$$CI_j(\vec{C}, \{i\}) = \frac{C_{max_j}(\vec{C}, \{i\}) - C_{min_j}(\vec{C}, \{i\})}{absmax_j - absmin_j} \tag{15}$$

The explanation model CIU calculates with the function CI_j the importance of feature i in the feature vector $\vec{C}$ for an output label (value) j. The function C_{max_j} determines the maximum output of the prediction j for a certain feature i. The C_{min_j} calculation follows a similar approach. The functions $absmax_j$ and $absmin_j$ determine the highest and lowest prediction from the given data set. More details are provided in [19, 25, 26].

The CIU values are in the interval between 0 and 1, where a higher numerical value represents a greater influence on the prediction and vice versa. A value of 0 means that the attribute does not influence the prediction. In contrast to the other methods, in which negative values argue against a classification, this is not the case here (i.e., values are between 0 and 1).

The advantages of the CIU algorithm are the different working calculation model and the absence in the questions of consistency (such as in LIME), as well as the calculation time as in SHAP [1]. Thus, this algorithm is suitable as a control method in the group of model-agnostic methods like in [9]. In [1] it is shown that CIU has the best stability during different runs with the same instances to explain, i.e. in terms of repeatability it is a good benchmark for SHAP and LIME. The problem with the value interval of 0 to 1 was solved in [9] by introducing a threshold value as a decision boundary. If feature values exceed this limit, the value has a positive influence and vica versa. Despite the advantages of the CIU algorithm, it is also non-deterministic, due to its randomized data sampling for numeric values, which the HR use case currently is [25].

3 The Idea

According to [2] there are two key criteria when you want to quantify the stability of explanations. The first aspect is the *Identity* criteria, i.e., that you should have identical explanations for identical instances. The second criterion is the *Separability*, which means that when the data instances are not identical, the explanation should not be identical.

The initial proposal of this work is to understand that these two criteria are at different ends of the same scale. The next step is to expand this scale by the criteria that are also presented in subsubsection 2.1.1. A representation of the so-called *Stability Scale* is shown in Figure 3. The following sub-sections explain the development of the scale in greater detail.

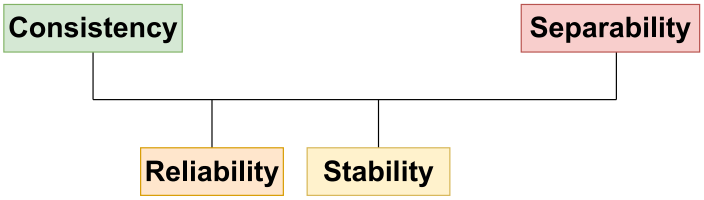


Figure 3: Stability Scale and their aspects (sub-criteria, properties)

3.1 Unification of Notions

For the standardization of notions (terms) the left end of the scale in Figure 3, serves as a fix point, from that other properties (terms), or criteria are explained. In Figure 2 there are two notions shown at this fix point: *Consistency* and *Identity*. According to the definition of *Identity* in [2] it is mainly focused on the instances and this could lead to the idea that only the data instances are mentioned. In contrast, the definition of *Consistency* by [4] includes the data instances and the predictor is considered. That means, the scope is broader and not only data-centered. Due to this, the notion *Consistency* is preferred, because of its wider scope.

Since the term *Reliability*, which comes next to *Consistency*, is not considered in the other works, it is not necessary to unify this notion with another.

Stability comes next to *Reliability* and in Figure 2 it is shown that there are three notions. The definitions of the notions are very similar, according to the majority usage of the term *Stability*, it could be used as a common notion.

The right end of the scale, which is denoted with *Separability*, is only mentioned by [2] and there is also no need for a standardization.

It could be summarized that the initial idea, denoted in this work as *Explanation-Stability*, could be seen as an aspect of an explanation produced by an XAI algorithm. When it comes to details, it could be stated that this aspect of stability has sub-criteria and the initial notion is not sufficient, that means the *Stability Scale* represents all kinds of aspects related to the proposed notion *Explanation-Stability*.

3.2 The Notions of this work in Detail

In the following, it is explained why the notions (sub-criteria) are ordered as seen in Figure 3. As a foundation, the arrangement of the *sub-criteria of Explanation Stability* in Table 5 was employed.

Table 5: Comparison of sub-criteria from the Explanation Stability according to [1, 2, 3, 4, 5]. The abbreviations used in the table are: id. as *identical*, sim as *similar*. The aspects of the comparison are the input, the output (that is, the explanation), and the constraints. The yellow-highlighted terms indicate the differences from the previous row, specifically in comparison to the stricter criterion.

Notion	**Inputs**		**Explanation**	**Constraints**
	model	data		
Consistency	id.	id.	id.	determinism
Reliability	id.	id.	highly sim.	–
Stability	id.	sim.	sim.	–
Separability	id.	non-id.	non-id.	different instances

The sub-criteria *Consistency* is the most strict criterion because of the common definitions of [2] and [4]: "identical inputs should result in identical outputs". Despite this, [3] mentioned in their definition of *Stability* that: "Nonetheless, a lack of stability can also be created by non-deterministic components of the explanation method, such as a data sampling step. Regardless of that, high stability is always desirable". In this work, where *Stability* is examined in detail, the constraint of determinism is assigned to the *Consistency* criteria as the strongest requirement for stability.

A weaker requirement in terms of stability is *Reliability*. In Table 5, the second row summarizes this criterion. The inputs are the same as in the *Consistency* row, with the explanation being only *highly similar* compared to the first criterion. Therefore, this stability aspect could be fulfilled by nondeterministic XAI algorithms. To ensure uniform inputs and solely vary the output – specifically, the explanation property – the stability aspect here is weaker compared to *Consistency*.

Another stability property denoted by *Stability* is not as strong as the *Reliability* criterion. The only identical input is the model (i.e., the predictor), while the data instances are only similar, representing slight variations of an original data point. Consequently, the explanations are only *similar*, not *identical* or *highly similar*. This implies that the stability aspect, denoted as *Stability*, has less stringent requirements than both *Consistency* and *Reliability*. Consequently, it can be concluded that *Stability* is the least demanding stability criterion represented in Figure 3.

The *Separability* property, situated at the opposite end of the scale, asserts that two fundamentally different instances should not produce the same explanation on the same ML model. This high instability contrasts with or complements *Consistency*, measuring an opposite aspect on the *Stability Scale*. The summarized information is visualized in Figure 4.

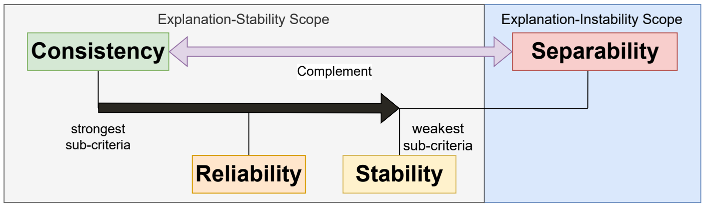


Figure 4: The proposed *Stability Scale* is shown here with additional context, starting from the left with the strongest criterion, up to the right with the complement of stability. The relations between the stability aspects: *Consistency, Reliability and Stability* are presented in this overview.

3.3 Mapping of Stability Criteria and Measures

At this point, it can be stated that there is an overarching concept for the stability of an explanation produced by an XAI approach. This term was introduced as *Explanation-Stability* and has three sub-criteria of stability. In order to quantify the *Explanation-Stability* objectively and to take all aspects into account, one should measure all sub-criteria of stability.

However, the properties of the XAI algorithms should be taken into account (subsection 2.3). In the proposed use case of [1], all XAI procedures are post-hoc approaches that generate local explanations. The explanation is of the type where each predicted feature $f(x_i)$ is associated with a feature importance value $\Phi(f, x)_i$. Therefore, for every explanation method Φ, the stability has been measured.

In case that all XAI approaches lack in terms of determinism, this stability criterion is very hard to fulfill by the used approaches. Therefore, more realistic stability measures such as *Reliability* and *Stability* are focused when measuring nondeterministic XAI algorithms. An overview of the measurement approaches assigned to the XAI methods used is shown in Table 6.

After assigning concrete stability measures, it is now denoted as *Explanation-Stability-Metric* due to the use of single stability-metric approaches.

Table 6: Assigning the proposed stability measures to non-deterministic, post-hoc explanation methods is discussed in detail in section 2.

	Explanation-Stability		
XAI Algorithm	**Consistency**	**Reliability**	**Stability**
LIME (2.3.1)	violate determinism	R(x)	S(x)
SHAP (2.3.2)			
CIU (2.3.3)			

3.4 Problems of Reliability and Stability

Even though the proposed approaches of the reliability measurement, presented in [1], as well as the stability analysis in [5] deliver highly valuable information about the quality of a certain XAI algorithm, the information are limited to one sub-criteria of the explainable models. The confidence interval method in [1] focuses on approaching the problem of *Reliability*, by conducting a test with the same data samples/models and with the expectation of highly similar data explanations. However, this approach lacks information regarding the stability of an XAI model.

To solve that, the paper [5] implements a method to evaluate the stability of an XAI model, by conducting changes on the input data. In contrast to [1], however, information regarding the reliability are missing. In addition, outliers are treated as regular explanations, which negatively effects the overall stability performance of the algorithm. An example for that can be seen in Figure 5, where several XAI algorithms are compared based on the similarity of the explanations in the feature "Income". By that, it can be seen that for instance LIME suffer from a high data dispersion because of several outliers. Due to a 1% change in one of the input features, the explanation similarity is expected to be in a range of 99% to 101%, where 100% refers to an identical explanation. However, algorithms like LIME seem to be prone to outliers. To avoid an overall biased score, those outliers needs to be handled accordingly.

Besides those problems, both papers do not provide a reference point, specifically an optimality criterion that serves as a baseline for gauging the algorithm's performance at its optimum. Instead, the only available approach is to compare algorithm A with another XAI method to determine whether algorithm A outperforms or underperforms in relation to algorithm B.

It can be concluded that stability measures are highly valuable when, combined because they address both stability aspects. Additionally, the use of an optimality criterion for the reliability metric is particularly interesting when evaluating the stability of an individual XAI algorithm. When considering the stability metric, it is important to take outliers into account.

3.5 Improvements of Reliability and Stability

To solve the problems mentioned in subsection 3.4, the current article presents a novel metric to evaluate current XAI solutions by combining the strengths of several XAI criteria. In particular, based on the approaches presented in equations $R(x)$ and Equation 12, the metric returns a vector of the corresponding scores to evaluate the XAI models. Since multiple aspects are considered, the metric provides an objective overall perspective on the performance of an XAI algorithm. In addition to that, the result vector enables comparability to other XAI models to find the best solution for the corresponding use-case.

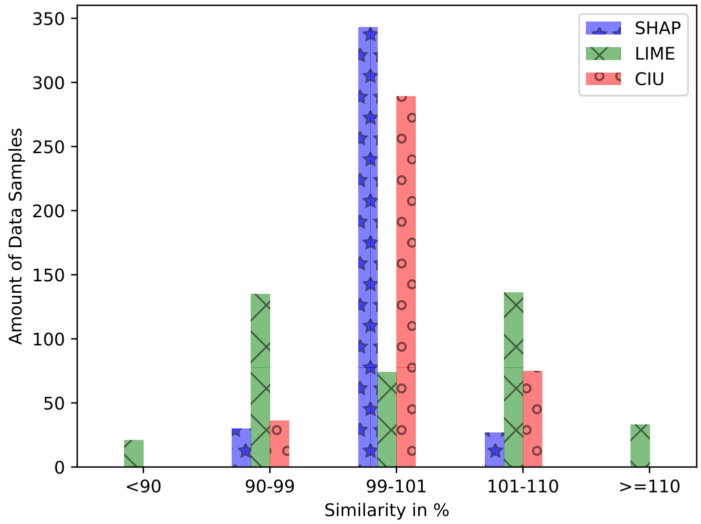


Figure 5: Algorithm Comparison about the similarity of the explanations for the feature „Income" - representation of the feature importance similarities for the data pairs. The similarity measures the percentage match between the explanation of the original data sample as well as the one with a feature change of around 1%. The results of *SHAP*, *LIME* and *CIU* are compared.

4 Results

The motivation for the novel approach to assess the stability aspects of XAI explanations was derived from the need to unify various notions. A comprehensive foundation was established through an extensive review of survey papers and articles on stability, including references such as [1]–[5]. This research aided in the identification and consolidation of the most relevant concepts and terms, forming the basis for the proposed *Stability Scale*, as depicted in Figure 2.

After gathering all key ideas and notions, a suitable process for unifying terms is detailed in subsection 3.1. The proposed and consistent scale is presented in Figure 3. The suggested notion of the XAI explanation stability is *Explanation-Stability*, it is important to emphasize that this is an overarching concept. For a more detailed view, the *Explanation Stability* is defined by three sub-criteria. These are visualized on a scale and ordered by their strictness according to Table 5. In fact, by defining the table and identifying these sub-criteria, research question RQ1 is fully answered.

Considering the real-world scenario presented in [1], it can be concluded that the *Explanation-Stability* construct of nondeterministic XAI approaches deteriorates in only two of the three criteria, namely *Reliability* and *Stability*. Furthermore, the outstanding idea of this work is to propose a concrete measurement approach for every stability aspect (sub-criterion). The sub-criterion of *Reliability* is measured by $R(x)$ and *Stability* is measured by $S(x)$, this concept is denoted as *Explanation-Stability-Metric*. Measures are described in depth in subsubsection 2.1.2 and improvements are identified in subsection 3.4.

This level of explicitness is unprecedented; unlike [2, 3], which provided only evaluation criteria and definitions. These works and considerations were pivotal when addressing the overarching question: “How can we evaluate XAI in general?” When the focus narrows to evaluating specific aspects, such as the actual explanation, and the scope is tightly defined, sources are rarely limited. Developing a new standard for evaluating XAI methods entails constructing a robust foundation from an initial and intuitive idea. This process involves being highly specific and thereby bridging existing gaps.

In addressing research question RQ2, the newly proposed approach, known as *Explanation-Stability-Metric*, has the potential to provide at least a partial answer to nondeterministic XAI approaches. With additional effort, expanding the metric to incorporate the stability aspect of *Consistency*, it could also be applied to deterministic XAI methods.

The key idea is to provide a holistic measure of *Explanation-Stability*; therefore, it is necessary to furnish information about stability. Upon the development of an initial outcome from the execution of this metric, a stability vector would provide all stakeholders with sufficient information to better trust, build, and comprehend AI models. Furthermore, this work should contribute to making autonomous systems applicable in high-risk domains.

5 Discussion

Future work and assumptions are considered to enhance the new stability assessment approach outlined in this study. Firstly, the question arises: “How useful is the addition of the stability metric for *Consistency* concerning nondeterministic XAI methods?”. This can be achieved by incorporating a distance measure that supports multidimensional data, such as the Euclidean distance. This enhancement provides a more objective overview of stability, which is valuable even when these methods produce less accurate scores than deterministic methods.

In the proposed stability metric, a complementary aspect to *Consistency*, referred to as *Separability*, was identified. Its addition could prove beneficial not only for measuring stability aspects but also for assessing *instability*, as mentioned in subsection 3.2. This broadens the scope of stability and introduces a related aspect into the metric. However, the inclusion of these two additional criteria requires further exploration in future work.

Considering the scope of stability, measuring all four sub-criteria could prove valuable, potentially making this approach applicable to deterministic XAI algorithms.

When considering the representation of the results, a second significant question arises: how to combine the measures of all sub-criteria into a single score. Several reasons justify this additional processing step. The fundamental concept of a metric is to distill relevant information into a real-valued single score. This approach offers a significant advantage, preventing information overload for those seeking insights.

As AI systems become more prevalent in our daily lives, ensuring information about stability is accessible to non-experts is essential. A single-score approach facilitates this by offering a concise and easily understandable representation.

However, aggregating the information also introduces a few disadvantages as it may lead to the loss of detailed information when reducing the sub-criteria scores of *Explanation-Stability* into a single score. When the resulting scores from each sub-criterion are condensed into a single value, questions such as “Could *Stability sub-criterion A* potentially compensate for the lower performance of *Stability sub-criterion B*, and to what extent does this influence exist?” arise.

Other important considerations include “How strong is the correlation between these stability aspects?”, “Is there a dominant stability criterion?” and “Are there domain-dependent sub-criteria that should be prioritized, such as in the medical field or other sensitive use cases?” These are the research questions that will need to be explored as this metric approach is implemented in real-world use cases.

6 Conclusion

In conclusion, the development of autonomous systems must adhere to specific quality criteria. Given that these systems frequently employ artificial intelligence for critical tasks, assessing the stability of complex machine learning models is imperative for comprehension, creation, and improvement. This work aims to lay the groundwork for future endeavors in verifying the stability of such models and their applicability in real-world scenarios.

Conflict of Interest The authors declare no conflict of interest.

References

[1] F. Gawantka, A. Schulz, J. Lässig, F. Just, “SkillDB - An Evaluation on the stability of XAI algorithms for a HR decision support system and the legal context,” in 2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 183–190, 2022, doi: 10.1109/ICCICC57084.2022.10101657.

[2] H. Löfström, K. Hammar, U. Johansson, “A Meta Survey of Quality Evaluation Criteria in Explanation Methods,” in J. De Weerdt, A. Polyvyanyy, editors, Intelligent Information Systems, 55–63, Springer International Publishing, Cham, 2022.

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

[4] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, “From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai,” ACM Computing Surveys, **55**(13s), 1–42, 2023.

[5] F. Gawantka, F. Just, M. Ullrich, M. Savelyeva, J. Lässig, “Evaluation of XAI Methods in a FinTech Context,” in International Workshop on Artificial Intelligence and Pattern Recognition, 143–154, Springer, 2023.

[6] M. Schuld, F. Petruccione, Machine Learning with Quantum Computers, Quantum Science and Technology, Springer International Publishing, 2021.

[7] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach (4th Edition), Pearson, 2020.

[8] I. Kakogeorgiou, K. Karantzalos, “Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing,” International Journal of Applied Earth Observation and Geoinformation, **103**, 102520, 2021, doi:https://doi.org/10.1016/j.jag.2021.102520.

[9] S. Hariharan, R. Rejimol Robinson, R. R. Prasad, C. Thomas, N. Balakrishnan, “XAI for intrusion detection system: comparing explanations based on global and local scope,” Journal of Computer Virology and Hacking Techniques, 1–23, 2022.

[10] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, "On the (In)fidelity and Sensitivity of Explanations," in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019.

[11] D. Minh, H. X. Wang, Y. F. Li, T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," Artificial Intelligence Review, 1–66, 2021.

[12] V. Arya, R. K. E. Bellamy, P. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," CoRR, **abs/1909.03012**, 2019.

[13] Q. V. Liao, M. Singh, Y. Zhang, R. Bellamy, "Introduction to explainable AI," in Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1–3, 2021.

[14] A. Das, P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," 2020, doi:10.48550/ARXIV.2006.11371.

[15] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in 2022 ACM Conference on Fairness, Accountability, and Transparency, 2239–2250, 2022.

[16] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," Advances in neural information processing systems, **31**, 2018.

[17] M. T. Ribeiro, S. Singh, C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 1135–1144, Association for Computing Machinery, New York, NY, USA, 2016, doi:10.1145/2939672.2939778.

[18] S. M. Lundberg, S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, 4768–4777, Curran Associates Inc., Red Hook, NY, USA, 2017.

[19] K. Främling, "Decision theory meets explainable AI," in Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2, 57–74, Springer, 2020.

[20] R. W. Floyd, "Nondeterministic algorithms," Journal of the ACM (JACM), **14**(4), 636–644, 1967.

[21] M. Knap, "Model-Agnostic XAI Models: Benefits, Limitations and Research Directions," 2022.

[22] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, "Explainable AI methods-a brief overview," in International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, 13–38, Springer, 2020.

[23] C. Molnar, "Interpretable machine learning," https://christophm.github.io/interpretable-ml-book/, 2022, accessed on 04 16, 2023.

[24] R. Heese, S. Mücke, M. Jakobs, T. Gerlach, N. Piatkowski, "Shapley Values with Uncertain Value Functions," in International Symposium on Intelligent Data Analysis, 156–168, Springer, 2023.

[25] K. Främling, "Contextual Importance and Utility: A Theoretical Foundation," in G. Long, X. Yu, S. Wang, editors, AI 2021: Advances in Artificial Intelligence, 117–128, Springer International Publishing, Cham, 2022.

[26] K. Främling, "Explainable AI without Interpretable Model," CoRR, **abs/2009.13996**, 2020.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).