

# Quantifying Explanation Stability and Robustness

Jero Schäfer<sup>1</sup>, Jakob Vanek<sup>1</sup> and Lena Wiese<sup>1</sup>

<sup>1</sup>*Institute of Computer Science, Goethe University, Robert-Mayer-Str. 10, 60325 Frankfurt am Main*

## Abstract

The increasing complexity and application of Machine Learning (ML) models necessitate the need for interpretable and reliable explanations to support decision-making, particularly in critical fields such as healthcare. Explainable Artificial Intelligence (XAI) techniques help to gain trust by providing explanations to the opaque black-box models. However, it is challenging to find the best methods bearing high explainability amongst multiple approaches. In this paper, we present a novel approach to quantifying the explainability of ML models, focusing on the key characteristics of stability and robustness. Our framework is demonstrated using a mortality prediction dataset for stroke patients, highlighting the practical implications of our approach and proving its general applicability. This research contributes to the field of XAI by providing a comprehensive, automated evaluation of model explainability facilitating the model life cycle and decision-making processes for different settings.

## Keywords

Explainable AI, Explainability, Machine Learning, BI, Stability, Robustness

## 1. Introduction

The growing power of ML models manifests in an increasing number of applications of such models in a variety of modern systems. Through their support humans can solve complex problems more efficiently and accurately. However, human subjects are still responsible for the decision for specific actions to be taken based on the model outputs and need to understand the behavior and reasoning behind the prediction. Unfortunately, this is not always the case and complex problems, in particular, often require incomprehensible but powerful models (black-box). The generation of explanations for a certain output enables more sound decision making and prevents blindly following the machine's suggestion.

Finding the optimal decision is crucial in fields such as medicine where the health of individuals is in threat. Especially, the prediction of mortality in conjunction with severe diagnoses (e.g., heart failures or strokes) can be life-saving if the risk of death is recognized and treated accordingly. For this it is extremely helpful if doctors are not only presented with the prediction of a model but additionally have access to explanations facilitating their evaluation of the situation. As not every model is well explainable, there is a growing interest in the assessment of the explainability of models. Ideally, expressive explainability measures enable model selection and estimation of the reliability of the explanations to improve the decision making.


---


*LWDA'24: Lernen, Wissen, Daten, Analysen. September 23–25, 2024, Würzburg, Germany*

✉ [jeschaef@cs.uni-frankfurt.de](mailto:jeschaef@cs.uni-frankfurt.de) (J. Schäfer); [vanek.jakob@gmail.com](mailto:vanek.jakob@gmail.com) (J. Vanek); [lwiese@cs.uni-frankfurt.de](mailto:lwiese@cs.uni-frankfurt.de) (L. Wiese)

🌐 <http://www.dbda.cs.uni-frankfurt.de/> (L. Wiese)

🆔 0000-0001-7727-1181 (J. Schäfer); 0000-0003-3515-9209 (L. Wiese)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In our work, we investigate on the automated labelling of ML models for the abstract criteria of explainability. To this end, we propose an explainability quantification based on the two characteristics stability and robustness. For each, a general, model-, explanation type-, and data-agnostic formula is presented that allows to quantify explanation stability and robustness for various scenarios with an intuitive value. We assemble a mortality prediction dataset for a cohort of stroke patients, which we analyze regarding the benchmark of the proposed explainability quantification approach. Furthermore, we also benchmark our measures for a classic visual classification task of handwritten digits. The rest of this work is organized as follows: Section 2 briefly illustrates recent approaches similar to our research, which is presented in Section 4. The cohort selection for our stroke dataset, the preprocessing, and classification model are described in Section 5. The results of the explainability benchmarks are discussed in Section 6. A summary and an outlook into future research directions are given in Section 8.

## 2. Related Work

There has been an uprising interest in the field of XAI in the past decade. Explaining the outputs of a black-box model trained with ML techniques to predict the class label of an unseen input is quite valuable in different scenarios. Adadi et al. [1] describe justification, control, improvement and discovery as reasons for generating explanations for black-box model predictions. These important factors induce benefits like transparency for the model's behavior or the potential for interactivity or refinement. Next to their predictive capabilities, the generation of explanations supports the trust in ML models [2] and yields important insights that aid decision making. In their review, Vilone et al. [3] give a systematic overview on publications using notions that have some relation to explainability, XAI methods and evaluation approaches. Their study hierarchically categorizes evaluation metrics for the different explainability attributes as objective automated or human-centered manual approaches.

Model Cards [4] or Data Sheets [5] formally describe AI models and training data to assess their applicability in other use cases and ideally certify of trained models. Seifert et al. [6] propose a framework for creating consumer-friendly labels for machine learning models to enhance transparency and trust. It introduces a method to summarize complex model characteristics into understandable labels, akin to nutrition labels on food products. The authors outline the design, evaluation, and implementation of these labels, emphasizing their potential to inform users about model performance, data usage, and ethical considerations. Nauta et al. [7] provide a comprehensive overview of the evaluation methodologies for XAI highlighting the shift from anecdotal and qualitative assessments towards more rigorous and quantitative evaluation frameworks. The authors categorize existing evaluation methods as Co-12 properties and identify gaps, emphasizing the need for standardized and reproducible metrics. The paper concludes by proposing a set of best practices and future directions to enhance the robustness and reliability of XAI evaluations. Chmielinski et al. [8] introduce an enhanced version of the dataset nutrition label, aimed at improving transparency and understanding of datasets used in AI development. It emphasizes the importance of context in mitigating potential harms by providing detailed metadata about the datasets, including their origin, composition, and potential biases. The authors propose a standardized framework for dataset documentation, facilitating

better-informed decision-making in AI deployment. Liang et al. [9] provide a comprehensive analysis of AI model documentation by examining 32,000 AI model cards. The authors identify common practices, gaps, and inconsistencies in AI model documentation, showing the need for standardization and improved transparency. They categorize the documented information into various dimensions, such as intended use, performance metrics, and ethical considerations, and evaluate the prevalence and quality of these dimensions. Boggust et al. [10] introduce a comprehensive framework for evaluating and comparing saliency methods, which are techniques used to interpret machine learning models by highlighting important input features. The authors present “Saliency Cards”, a structured approach that categorizes these methods based on various dimensions such as methodology, output, and application context, facilitating a systematic comparison. The paper aims to provide insights into the strengths and weaknesses of different saliency methods. Tagliabue et al. [11] propose the DAG Card as an extension and enhancement of the Model Card concept. It introduces Directed Acyclic Graph (DAG) structures to represent complex AI models and their deployment scenarios comprehensively. The DAG Card aims to provide detailed explanations of model architecture, data processing steps, and performance metrics across different stages of a model’s lifecycle.

While these papers emphasize the importance of XAI, their focus lies on a descriptive assessment but not on a framework to quantify the explainability of AI models. Instead, a framework with visual components like an easy-to-use, web-based dashboard [12] presents compressed information about the model characteristics like its explainability. The quantification of model explainability reaches a broad audience (e.g., management, developer team, end users) as the offered information can be consumed easily. As a major benefit, the usage of such tools amplifies the integration of AI into workflows while increasing the understanding of and trust in AI.

Two interesting notions related to explainability are robustness [13, 14] and stability[15], which are investigated in this work and formal definitions are proposed to calculate metrics for the assessment of explainability based on these notions. Both suggested criteria fall under the Co-12 [7] concept of continuity, i.e., the generalization capability of an XAI method. However, all of the categorized continuity approaches fail to generalize as they are explicitly formulated for a certain classification model, explanation type, explanation distance, or data type. As solution to this, we instead aim for a general formulation of our metrics to be applicable to a wide variety of scenarios and not being dependent on specific properties.

### 3. XAI Business Use Cases

In general, our framework can handle XAI in several different business use cases:

- XAI in the classification of tabular data. For example, in the health care domain data often come as electronic health records (EHR) in a tabular format on which a medical decision support system can be based. Explanations for model predictions improve the outcome for the patient by providing additional information to the medical staff.
- XAI in visual classification where data come as images and have to be classified in an automated way, e.g., filtering faulty products in a production line or processing handwritten documents automatically for digital archiving.

- Text classification as a subpart of natural language processing (NLP) that can be used for automated customer support (like chat bots or automated response systems) or in the area of marketing (like social media analysis or continuous customer feedback analysis).

For each of these use cases, XAI methods can effectively improve the usage of self-service ML by supporting decision making and reasoning. The AI model lifecycle is enhanced by the additional information developers and stakeholders are presented with, e.g., in model debugging and selection or gaining insights and trust into the behavior of the black-box. Ideally, the XAI methods demonstrate capabilities of the model in a comprehensible fashion not only for ML experts but also end user or even experts of other fields like medicine or finance. The generated explanations help to improve processes by highlighting potential weaknesses or revealing critical failures. In the scope of this article, we apply our two proposed explainability metrics exemplary to two different scenarios. As pointed out in the previous section, our general quantification approach overcomes the limitation of previous works where only problem-specific computations were carried out. We benchmark our metrics for the explanations generated for two use cases: a stroke mortality tabular classifier and a handwritten digit visual classifier. The results demonstrate the practicality of our approach and enable judging and comparing explanation generation across XAI methods.

#### 4. Explainability of XAI Methods in Classification Problems

XAI comprises methods to make the outputs of AI systems understandable and interpretable to humans. From a Business Intelligence point of view, the deployment of AI tools makes it mandatory that users and stakeholders can comprehend how and why an AI system reaches certain conclusions or predictions. The application of XAI methods facilitates the interpretation of the decisions in business analytics settings and automates the generation of explanations. However, the quality of the presented explanations is to be assessed to draw justified conclusions.

There exist multiple definitions of model *explainability* or *interpretability* (often used synonymously) referring to the capability to extract “the meaning of an abstract concept” [3]. Explainability is related to and often built on top of several characteristics, e.g., transparency, faithfulness or understandability. A superior explainability is therefore bound to a high-quality adherence by the explanations of the model outputs to its building blocks. For example, a transparent model is per-definition more explainable than an opaque one and explainability requires the explanations to be understood by humans. Our research focuses on the as far as possible automated evaluation of explainability metrics for a tabular binary classifier such that no human interaction is required and the generation of a label for explainability is simple. To this end, we propose two novel, general-purpose measures for characteristics of explainability, namely stability and robustness, in the following.

We denote the explanation generation for instance  $x_i$  or – more precisely – the prediction of  $x_i$ 's class label,  $M(x_i)$ , through the classifier  $M$  by  $E_M(x_i)$ . The explanation generation mechanism  $E_M$  is required to support local explanations, e.g., a SHAP [16] or lime [17] tabular explainer. The expression  $E_M(\{x_1, \dots, x_p\})$  then refers to the set of explanations  $\{E_M(x_i) \mid i \in \{1, \dots, p\}\}$  generated for an input set of  $p$  instances. Based on an explainer  $E_M$  for a classifier  $M$ , an evaluation function  $f_W(E_M) \in [0, 1]$  measures a specific characteristic  $W$  of the explainer. An evaluation

value of  $f_W(E_M)$  close to 1 indicates a high performance of  $E_M$  regarding  $W$  and a value closer to 0 shows that  $W$  is satisfied only weakly or insufficiently. To summarize different characteristics, a single value is calculated as judgement of the capability of  $M$  of being well explainable by  $E_M$ , e.g., as a weighted sum  $\sum_{i=1}^z \omega_i \cdot f_{W_i}(E_M)$  over characteristics  $W_1, \dots, W_z$ .

#### 4.1. Stability of Explanations

Explanation generation is considered stable if similar instances with the same labels have similar local explanations [15]. The most direct way to calculate the stability is to find a set of neighboring instances  $N(x_i) = \{x_{j_1}, \dots, x_{j_k}\}$  for each instance  $x_i$  in (a subset of) the dataset, e.g., with a  $k$ -nearest neighbor graph. Then, the average similarity or distance between the explanation  $E_M(x_i)$  and the set of explanations  $E_M(N(x_i)) = \{E_M(x_{j_1}), \dots, E_M(x_{j_k})\}$  of the neighboring instances measure explanation stability. Our heuristic quantifies this property by the similarity of groups of explanations according to a grouping of the dataset instances. Therefore, clustering techniques divide the dataset into meaningful groups of similar instances in an unsupervised manner. Applied to both, two clusterings  $C_D$  and  $C_E$  are computed for the data in  $D$  and  $E = E_M(D)$ , respectively. The more similar the resulting clusterings are, the more stable are the explanations generated by  $E_M$ . In the following definition we extend the stability estimation by using sets of  $q$  clusterings for both  $D$  and  $E$ .

**Definition 4.1.** Given the two sets  $C^D = \{C_1^D, \dots, C_q^D\}$  and  $C^E = \{C_1^E, \dots, C_q^E\}$  of  $q$  clusterings of a dataset  $D$  and its explanations  $E = E_M(D)$ , respectively. The *stability* of the explanations generated by  $E_M$  is defined as the maximal pairwise clustering similarity:

$$f_{Stab}(E_M) = \max_{A \in C^D, B \in C^E} Sim(A, B) \quad (1)$$

The clustering similarity can be computed by selected external cluster validation indices (CVI) that operate on a clustering  $C$  given external information like the cluster labels from a reference clustering  $C'$ . The Rand Index (RI) therefore assesses  $Sim(C, C')$  by the ratio of pairs of objects that happen to be in the same or different clusters in both clusterings compared to all object pairs. Thus, the ratio yields the probability that  $C$  and  $C'$  partition a random selected pair of objects in an identical fashion. The Adjusted Rand Index (ARI) [18] overcomes the lack of a baseline in the RI. It corrects it for the chance by considering the expected agreement between both clusterings. It ranges in  $[-1, 1]$  where -1 refers to a completely different label matching, 0 to a random label matching, and 1 to a perfect label matching. Our approach allows to vary the clustering methods and adapt them to the dataset with a suitable distance metric.

#### 4.2. Robustness of Explanations

Robustness refers to the independence of marginal changes when processing inputs [13, 14]. As a consequence, marginal changes of data points not affecting the model's prediction should also lead to marginally divergent or identical explanations. This is closely related to the concept of stability with some variation in the level of granularity with which both approaches operate. Analogous to the stability, it could be measured using a more fine-grained clustering to average the intra-cluster explanation distances or a  $k$ -NN graph to aggregate the explanation distances

of neighboring data points with the same label for a small  $k$ . However, it is not trivial to automate the clustering process to ensure a fine granularity and clusters mimicking subsets of marginally different data points. This neighbor approach depends on the dataset structure and requires dense regions of data points to find enough neighbors. Instead, marginal differences of the data points can be sampled artificially from data points with minimal distance. This also emphasizes more on the distinction of robustness and stability (synthetic data vs neighboring instances). Hence, the explanation distances for the sampled data points and their origin measure the robustness. The smaller these distances are, the more robust the explanations are.

**Definition 4.2.** Given a subset size  $s$  and a number of samples  $k$ , the *robustness* of the explanations generated by  $E_M$  is measured as

$$f_{Robu}(E_M) = 1 - \frac{1}{s \cdot k} \sum_{x \in X} \sum_{y \in Q_k(x)} \frac{d(E_M(x), E_M(y))}{\max_{x', y' \in D} d(E_M(x'), E_M(y'))} \quad (2)$$

where  $X \subseteq D$  is a subset of  $|X| = s$  data points,  $Q_k(x)$  is a set of  $k$  marginally changed samples of  $x$  with  $M(x) = M(y)$ ,  $y \in Q(x)$ , and  $d(\cdot, \cdot) \geq 0$  is the pairwise explanation distance.

The robustness (Eq. 2) is estimated using a random subset  $X \subseteq D$  of  $s$  data points from  $D$ . For each data point  $x \in X$ ,  $k$  marginally different samples  $y \in Q_k(x)$  are generated. The distances between the explanations  $E_M(y)$  of the samples  $y$  and the explanation  $E_M(x)$  for the sample origin  $x$  are normalized by the maximum distance and averaged over all  $s$  data points of the subset  $X$ . The normalization provides the convenience of a more interpretable robustness value  $f_{Robu}(E_M) \in [0, 1]$  where 1 indicates robust explanations (that still have high similarity) and 0 refers to the explanation generation being affected by marginal changes (low similarity between the explanations of marginally different inputs). Actually, the robustness might also be slightly negative in specific cases due to the normalization constant. However, depending on the sampling mechanism  $Q_k(x)$ , there will be only small distances between original and sampled explanations such that quotient will be smaller than 1. It should always be checked that the model prediction is unchanged for a slightly perturbed input [19]. Enforcing the same predicted label  $M(x) = M(y)$  of the samples in Def. 4.2 also decreases explanation distances in comparison to the distance for a pair  $x', y' \in D$  with different predicted labels most of the time.

## 5. Experimental Setup

### 5.1. Feature Extraction and Preprocessing

Even though people tend to rely on AI to obtain useful insights without further guidance, usually still a certain degree of feature extraction and data preprocessing are needed. In fact, these steps ensure that the data used to train AI models is correct and relevant for the intended purpose, which in turn again enhances the reliability and interpretability of the AI outputs. Useful techniques that should be applied to data before AI training comprise data cleaning (removing noise, errors, and inconsistencies that may only be identified by a human expert); missing values handling (where the usefulness of techniques like proper imputation may be a question that can only be answered by humans) and standardization and normalization (where

appropriate techniques should be carefully chosen by humans depending on the use case). Hence, preprocessed data may lead to cleaner and more understandable models. This makes it easier to apply XAI techniques, which rely on the quality of the underlying data to provide meaningful explanations. More technically, in terms of feature importance, XAI methods like SHAP or LIME can more accurately determine the importance of different features if they are exposed to a manually determined subset of the entire feature set; dimensionality reduction itself can simplify the model, making it easier to interpret without any performance impact. Moreover, data preprocessing may identify and reduce any biases contained in the data, which is essential for XAI to provide fair and unbiased explanations.

For the experiments regarding explanation stability and robustness, we generated a real-world dataset from medical data obtained from the MIMIC-III (v1.4) [20] database. The database comprises hourly measurements from 58k hospital admissions of 45k de-identified patients in ICU in the Beth Israel Deaconess Medical Center, Boston, from 2001-2012. A multitude of features was measured per patient on an hourly basis in over 58k hospital admissions of 45k de-identified patients in ICU and successfully implemented in previous research [21, 22]; MIMIC-III has already been successfully implemented in previous research papers and was the basis for public PhysioNet challenges<sup>1</sup>. Thus, this dataset presents a reliable source with sufficient size for ML. The dataset was imported into PostgreSQL and necessary database setup files for this were distributed by PhysioNet<sup>2</sup> (detailed setup process in [23]). Amongst the admissions, we exclude underage patients or records with missing data and keep only patients with “chart\_events”-data (cf. [24]). Moreover, duplicated patients with multiple ICU transfers within a single hospital stay are removed. Based on previous research [25], all ICU-stays shorter than 24h are filtered out due to insufficient data resulting in 32k admissions. Often one would also filter for a specific hospital database system (“CareVue” vs. “metavision”) as it is challenging to map roughly 12k CareVue variables, encompassing vital signs and lab events, to 2k metavision variables and integrating both systems is a major task in itself. Thus, only selected chart\_events are mapped for our use case of stroke mortality prediction, combining hemorrhagic and ischemic cases. Cohort selection is based on the ICD9-codes of the diagnosis at admission: (1) hemorrhagic, (2) ischemic, and (3) other stroke types (incl. late effects). For further analysis, we pre-select 40 relevant features including patient demographics and their diagnosis. The patient vitals and lab results are selected as guided by medical expertise from previous research. One feature of note is the OASIS Score<sup>3</sup>, which represents a common and insightful index for patient mortality risk and helps comparing new prediction models with the status quo. The calculation of the score depends on multiple factors (ventilation, pre-ICU length of stay, etc.).

The careful preprocessing and filtering of admissions and measurements<sup>4</sup> leaves 9 mixed features for mortality prediction (e.g., OASIS score) including gender and ethnicity for 2655 unique ICU stays with diagnosis of (1) hemorrhagic, (2) ischemic, or (3) other stroke types (incl. late effects). We trained an XGBOOST (eXtreme Gradient Boosting) classifier to predict mortality of patients with such diagnosis (AUROC: 0.894, Accuracy: 0.868, Recall: 0.653, Precision: 0.511). The prediction results are not as performant as desirable but we refrained from further

---

<sup>1</sup><https://physionet.org/about/challenge/moody-challenge>

<sup>2</sup><https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iii>

<sup>3</sup><https://github.com/caisr-hh/Dayly-SAPS-III-and-OASIS-scores-for-MIMIC-III>

<sup>4</sup>Details under <https://github.com/jeschaef/explainability>

optimizations as our analysis focuses on quantifying explanation stability and robustness.

## 5.2. Tabular Classification

For the explainability evaluation of the model on the stroke cohort, we computed the stability using three different clustering algorithms: k-Means, hierarchical and spectral clustering. The number of clusters  $k$  can be controlled in each of them allowing for a straightforward creation of  $q$  different clusterings by scaling  $k$ . We employed the ARI, Adjusted Mutual Information (AMI) and V-Measure (V) to find the maximum similarity of a pair of clusterings for assessing stability. The explainers for the generation of explanations of the tabular classifications are SHAP [16], lime [17], Anchor [26] and Accumulated Local Effects (ALE) [27]. The chosen distance metrics are the common Euclidean distance  $\|\cdot\|_2$  for SHAP, lime and ALE, and the Jaccard distance on the subsets of data points satisfying the anchors. The sampling function  $Q_k(x)$  uses the Signal-To-Noise-Ratio (SNR) as additional parameter scaling the perturbation added to  $x$ .

## 5.3. Visual Classification

The classification of images for certain task is another popular field of ML. There exist various examples proving the strength of visual classification models in comparison to human evaluation. For example, this can be applied to diagnose with medical images like ECG or CT measurements or to detect faulty objects in production lines. In addition to powerful models, it is often helpful being able to gather additional information to the model’s prediction. The visual classification models are then usually explained by XAI methods that highlight certain areas of interest in the images, e.g., heatmaps for feature attribution, activation or saliency.

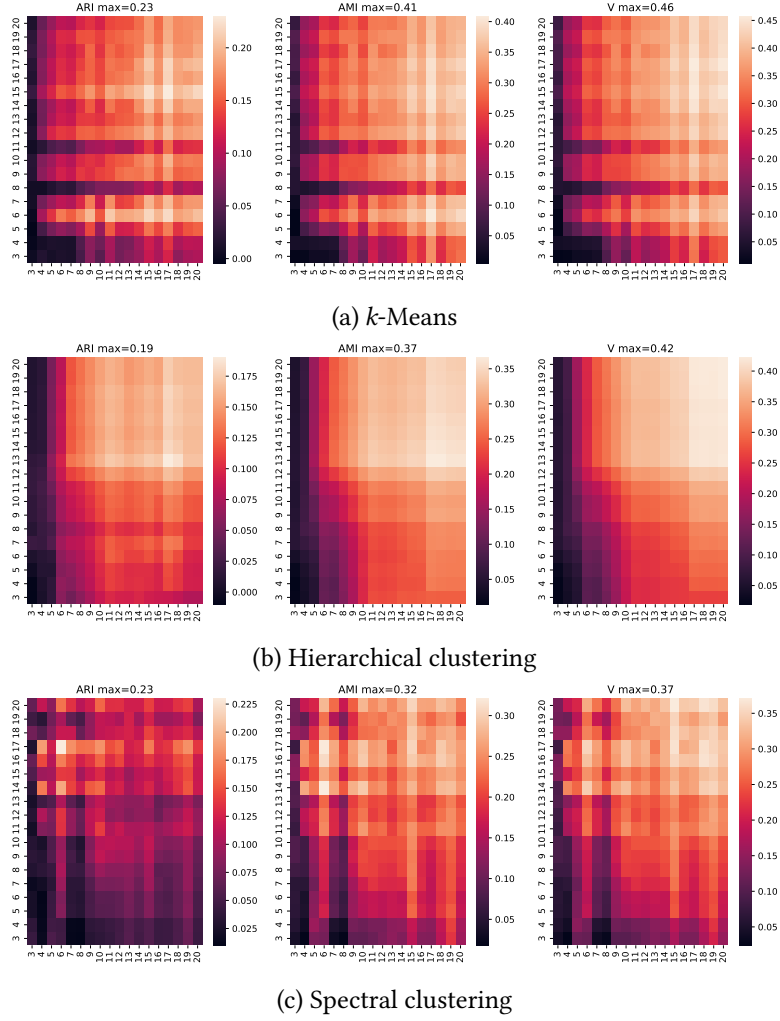
We apply our proposed explainability metrics for the example of the classification of handwritten digits with the famous MNIST dataset. This demonstrates the applicability of our measures for visual classification, which is different from the previously discussed tabular classification with regards to the type of data, the generated explanations and their respective distances. To this end, we train a basic CNN classification model by finetuning a pretrained VGG16 model to distinguish images of written zeros and ones and generate explanations in form of heatmaps using Grad-CAM [28] and Occlusion Sensitivity (OS). The heatmap distances are calculated using some common image similarity methods: Mean-Squared Error (MSE), Cosine-Similarity, and the Structural Similarity Index Measure (SSIM). The cosine similarity is computed for the flattened heatmaps. The sampling method for the robustness assessment simply adds gaussian noise (controlled by SNR) to a given image to perturb it and generate new sample images. To assess stability, distance matrices are precomputed for the pairwise heatmap distances and clustered hierarchically (SLINK). The cluster similarities indices are the same as before.

# 6. Results

## 6.1. Stroke Mortality Prediction

In Fig. 1, the stability results of the SHAP explainer are shown in form of heatmaps for the pairwise clustering similarities where the subfigures (a)–(c) each show three heatmap plots

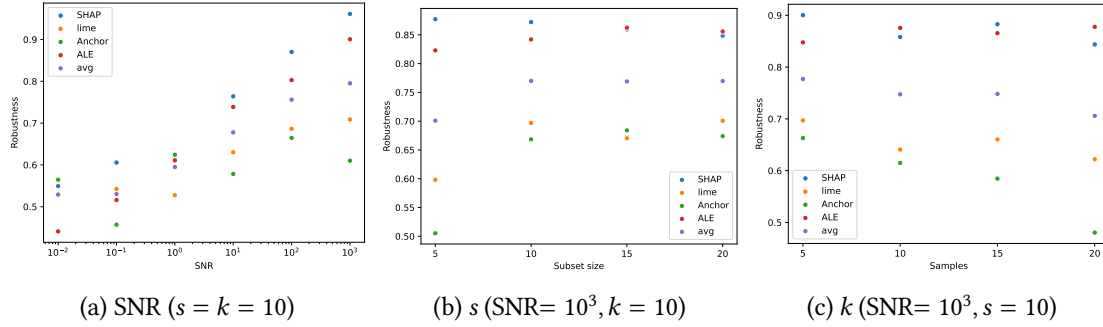




**Figure 1:** Stability  $f_{Stab}(E_{MM})$  of the SHAP explanations of the XGBoost classifier. The x- and y-axes show the number of clusters of the dataset and explanation clustering, respectively. The color brightness indicates the height of the similarity value according to the scale next to each plot.

for each similarity for a single clustering algorithm with  $k \in [3, 20]$ . All plots reveal a slight tendency towards pairs of clusterings with more clusters. For each clustering algorithm, one can also notice that the heatmaps are quite identical looking despite some minor deviations and the same patterns can be found more or less in the heatmap of each CVI. The main differences are the scales of the respective CVI, e.g., the maximal values were 0.23 (ARI), 0.41 (AMI) and 0.46 (V) with k-Means (Fig. 1a).

The smooth transitions of the “heat” in the plots of the hierarchical clusterings are caused by the marginal changes between clusterings with different numbers of clusters  $k$  and just different heights in the same underlying dendrogram. The maximum ARI values of the three clustering methods are also very similar ( $\sigma^2 \approx 10^{-4}$ ) whereas the maximum AMI and V values



**Figure 2:** Robustness  $f_{Robu}(E_M)$  benchmark with different parameters being scaled. The parameter value is mapped on the x-axis and the robustness on the y-axis. The colors indicate the different explainers (or average of all explainers).

diverge slightly more ( $\sigma^2 \approx 10^{-3}$ ) but show the highest similarity value for k-Means clusterings and the lowest for spectral clusterings. The results for the other explainers reveal less stable explanation generation with lime and ALE (max. similarity values 0.06-0.21 and 0.09-0.2) and a poor performance of Anchor (max. similarity values 0.0-0.1) due to unsuitable clusterings of the explanations based on the Jaccard distance.

The robustness evaluation of the four explainers on our classifier is shown in Fig. 2. The results of the individual scaling the three different parameters of  $f_{Robu}$  while fixing the other parameters is displayed in its three subfigures (a)-(c). The  $SNR$  has the main influence on the result of robustness as it controls the distance of the sampled data points (Figure 2a). With an increasing  $SNR$  also the estimated robustness increases as less perturbation is added in the sampling leading to less distance. The subset size  $s$  (Fig. 2b) and the number of samples  $k$  (Fig. 2c) do not scale the robustness at all (even for  $s$  or  $k=5$ ) but just decrease the fluctuation of the resulting robustness. Overall,  $s$  and  $k$  might be dependent on the number of data points, thus, it might be required adapting them for other, larger dataset sizes. For the stroke dataset,  $k, s \in \{5, 10\}$  already suffice for steady results and can be computed quicker. For  $SNR = 10^3$ , the SHAP and ALE explainer provide the most robust explanations ( $f_{Robu} \in [0.8, 0.9]$ ) and the lime and Anchor explainer perform slightly worse ( $f_{Robu} \in [0.5, 0.7]$ ).

In total, the explainability composed out of the two characteristics robustness and stability is 0.71 (SHAP), 0.46 (lime), 0.36 (Anchor) and 0.55 (ALE). Here, we assumed equal weights and the maximal robustness and stability result of each explainer. Even for the combination of only two characteristics it is hard to interpret this explainability value. Nevertheless, it summarizes the explanation generation performance and enables an automatic selection of the best performing methods regarding the functional measurement of continuity [7] with our two proposed formulas for stability and robustness.

## 6.2. Handwritten Digit Classification

The explanation robustness and stability results of the handwritten digit classification task with Grad-CAM and OS explanations are shown in Table 1 and 2, respectively. As expected, the Grad-CAM explainer shows an increasing explanation robustness with decreasing perturbation for

Explainer	MSE			Cosine			SSIM		
	$10^0$	$10^1$	$10^2$	$10^0$	$10^1$	$10^2$	$10^0$	$10^1$	$10^2$
Grad-CAM	0.33	0.50	0.69	0.82	0.81	0.82	0.50	0.73	0.91
OS	0.46	0.57	0.75	0.82	0.88	0.91	0.42	0.72	0.81

**Table 1**

Robustness evaluation ( $s = k = 10$ ) of the MNIST classifier for Grad-CAM and OS explanations using three distance metrics (MSE, Cosine, SSIM) for different  $SNR$ .

Explainer	MSE			Cosine			SSIM		
	ARI	AMI	V	ARI	AMI	V	ARI	AMI	V
Grad-CAM	0.98	0.95	0.95	0.03	0.03	0.09	0.0	0.0	0.07
OS	0.99	0.98	0.99	0.99	0.98	0.98	0.01	0.01	0.09

**Table 2**

Stability evaluation of the MNIST classifier for Grad-CAM and OS explanations using three distance matrices (MSE, Cosine, SSIM) and hierarchical clustering for  $k = 2, \dots, 20$ .

the MSE and SSIM. However, the robustness of the cosine similarity is constant for the different perturbation levels. The overall performance varies across the distance metrics with the best robustness value 0.91 for  $SNR = 10^2$  and SSIM. The other two distance metrics yield smaller values highest  $SNR$  level. We see a similar increasing trend for the OS explainer throughout the  $SNR$  scaling. In contrast to the Grad-CAM, the cosine-based value is the best performing. The values are comparable to those of the Grad-CAM approach revealing only slight differences.

We applied PCA dimensionality reduction to the embeddings of the input images from  $D$ . The resulting plot showed two clearly separable classes perfectly matching the ground-truth labels whereas the explanation clusterings differ in their performance regarding the baseline clustering. The stability results in Table 2 are in contrast to the robustness evaluation. The MSE-based clusterings indicate very stable explanations throughout all three CVIs and both explainers. For the Grad-CAM explainer, we see basically non-existent robustness near 0 with the cosine similarity and SSIM. This means only random cluster assignments of the explanations in comparison to the input image clusterings. The same happens for the OS explanations under the SSIM distance matrix, which seems to be not applicable for this approach. Still, the OS explainer has also almost perfect stability when applying the cosine similarity on the heatmaps. These results show a sensitivity towards distance metric selection, which must be performed with care to yield meaningful clusterings or otherwise random assignments.

## 7. Discussion

Despite the straightforward computation of stability and robustness, there are some aspects that need to be considered and investigated in future research. The first aspect is the choice of explanation distance metrics  $d(E_M(x), E_M(y))$  that determine the clustering used for the stability assessment as well as the robustness relying on these distances. For example, the experiments show that clusterings of Anchor explanations using the Jaccard distance on tabular classification

or the SSIM and OS explainer on visual classification are not similar to any of the clusterings of  $D$  itself and the generated explanations are quite unstable. Hence, it is crucial to carefully select an appropriate distance metric and clustering algorithm that capture the relationship between explanations such that it can produce meaningful groups and an adequate interpretation.

For both measures other factors need consideration when interpreting the results. As the parameters  $s$  and  $k$  (robustness) seem to not influence the results too much (if at least some randomness is introduced), the sampling algorithm might have more impact. In fact, there might be different results when applying various perturbations, e.g., image scaling, shifting, and rotation, or more sophisticated sampling. In future experiments, it could be checked if these actually provide more options to assess explanation robustness against manifold perturbations.

Another issue is the comparability of stability and robustness across different classifiers trained on the same data, the same classifier trained on similar datasets in the same scenario, or even both. Moreover, the relation to a specific explainer or explanation generation method has to be kept in mind as our experiments have shown for different methods. To preserve this variability, the proposed definitions of the stability and robustness characteristic are generic and allow for using different distance metrics or clustering algorithms. Therefore, there might be more suitable choices of explainers and parameters according to the model or dataset. If applicable, fixation of the dataset, classification model, clustering algorithm, and explanation distance metric at least provides facts about the stability and robustness of the explanations from similar explainers, e.g., multiple feature importance or heatmap explainers. This helps identifying the most useful explanations for the current scenario.

This flexibility of the stability definition also induces challenges towards the interpretability of the quantification of the clustering similarity. Romano et al. [29] recommend the usage of ARI for the case of big equal sized clusters in the reference clustering (here:  $C_i^D$ ) and AMI for an unbalanced clustering and small clusters. Furthermore, the impact of the choice of a random model for the adjustment of the clustering similarity measure and the comparison against a reference clustering or between two derived clusterings should be considered [30].

## 8. Conclusion

To make AI an acceptable tool in Business Intelligence and Analytics in particular – or in decision support tools in general – AI tools and frameworks have to achieve high levels of reliability, transparency and trust. These form the basis for AI accountability, especially in critical business applications (like healthcare). XAI methods provide explanations of how AI models arrive at specific conclusions, making it easier for business analysts and decision-makers to understand and trust these outputs – reducing the likelihood of misinterpretation and hopefully leading to more informed and effective business decisions. Quantifying the outcome of XAI methods is hence a paramount precondition to turn AI decisions into reliable business decisions. Our approach yields two general-purpose metrics to quantify stability and robustness of explanations. In future research, the applicability of various other explanation generation methods and corresponding distance metrics can be tested. The expansion to multiple other scenarios and classification tasks will show the generalizability of the explainability approach in different fields.

## References

- [1] A. Adadi, M. Berrada, Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence, *IEEE Access* 6 (2018) 52138–52160.
- [2] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, *International journal of human-computer studies* 58 (2003) 697–718.
- [3] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* 76 (2021) 89–106.
- [4] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.
- [5] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets, *Communications of the ACM* 64 (2021) 86–92.
- [6] C. Seifert, S. Scherzinger, L. Wiese, Towards generating consumer labels for machine learning models, in: *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, IEEE, 2019, pp. 173–179.
- [7] 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 (2023) 1–42.
- [8] K. S. Chmielinski, S. Newman, M. Taylor, J. Joseph, K. Thomas, J. Yurkofsky, Y. C. Qiu, The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence, *arXiv preprint arXiv:2201.03954* (2022).
- [9] W. Liang, N. Rajani, X. Yang, E. Ozoani, E. Wu, Y. Chen, D. S. Smith, J. Zou, What’s documented in ai? systematic analysis of 32k ai model cards, *arXiv preprint arXiv:2402.05160* (2024).
- [10] A. Boggust, H. Suresh, H. Strobelt, J. Gutttag, A. Satyanarayan, Saliency cards: A framework to characterize and compare saliency methods, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 285–296.
- [11] J. Tagliabue, V. Tuulos, C. Greco, V. Dave, Dag card is the new model card, *arXiv preprint arXiv:2110.13601* (2021).
- [12] J. Schäfer, L. Wiese, ASDF-Dashboard: Automated Subgroup Detection and Fairness Analysis, in: *Proceedings of the LWDA 2022 Workshops: FGWM, FGKD, and FGDB, Hildesheim (Germany), Oktober 5-7th, 2022*, volume 3341 of *CEUR Workshop Proceedings*, 2022, pp. 45–56.
- [13] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, *Visual Informatics* 1 (2017) 48–56.
- [14] D. Alvarez-Melis, T. S. Jaakkola, On the Robustness of Interpretability Methods, *arXiv preprint arXiv:1806.08049* (2018).
- [15] D. Alvarez-Melis, T. Jaakkola, Towards Robust Interpretability with Self-Explaining Neural Networks, in: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*. Montréal, Canada, volume 216, 2018.
- [16] S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett

- (Eds.), *Advances in Neural Information Processing Systems* 30, Curran Associates, Inc., 2017, pp. 4765–4774.
- [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*, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.
  - [18] L. Hubert, P. Arabie, Comparing partitions, *Journal of classification* 2 (1985) 193–218.
  - [19] W. Nie, Y. Zhang, A. Patel, A theoretical explanation for perplexing behaviors of backpropagation-based visualizations, in: *International conference on machine learning*, PMLR, 2018, pp. 3809–3818.
  - [20] A. Johnson, T. Pollard, R. Mark, MIMIC-III Clinical Database (version 1.4), 2016. doi:10.13026/C2XW26.
  - [21] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmarking deep learning models on large healthcare datasets, *Journal of biomedical informatics* 83 (2018) 112–134.
  - [22] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, K. Borgwardt, Early prediction of sepsis in the ICU using machine learning: a systematic review, *Frontiers in medicine* 8 (2021) 607952.
  - [23] A. E. W. Johnson, D. J. Stone, L. A. Celi, T. J. Pollard, The MIMIC Code Repository: enabling reproducibility in critical care research, *Journal of the American Medical Informatics Association* 25 (2018) 32–39.
  - [24] M. Moor, M. Horn, B. Rieck, D. Roqueiro, K. Borgwardt, Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping, in: *Machine Learning for Healthcare Conference*, PMLR, 2019, pp. 2–26.
  - [25] A. E. Johnson, J. Aboab, J. D. Raffa, T. J. Pollard, R. O. Deliberato, L. A. Celi, D. J. Stone, A comparative analysis of sepsis identification methods in an electronic database, *Critical care medicine* 46 (2018) 494.
  - [26] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-Precision Model-Agnostic Explanations, in: *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
  - [27] D. W. Apley, J. Zhu, Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82 (2020) 1059–1086.
  - [28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, *International journal of computer vision* 128 (2020) 336–359.
  - [29] S. Romano, N. X. Vinh, J. Bailey, K. Verspoor, Adjusting for Chance Clustering Comparison Measures, *The Journal of Machine Learning Research* 17 (2016) 4635–4666.
  - [30] A. J. Gates, Y.-Y. Ahn, The Impact of Random Models on Clustering Similarity, *arXiv preprint arXiv:1701.06508* (2017).