Riccardo Guidotti
Ute Schmid
Luca Longo (Eds.)

Communications in Computer and Information Science 2578

Explainable Artificial Intelligence

Third World Conference, xAI 2025
Istanbul, Turkey, July 9–11, 2025
Proceedings, Part III

Part 3

Springer

OPEN ACCESS

Communications in Computer and Information Science 2578

Rationale
The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics
The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors
Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process
The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing
CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start
To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.

Riccardo Guidotti · Ute Schmid · Luca Longo
Editors

Explainable Artificial Intelligence

Third World Conference, xAI 2025
Istanbul, Turkey, July 9–11, 2025
Proceedings, Part III

Springer

Editors
Riccardo Guidotti
University of Pisa
Pisa, Italy

Ute Schmid
University of Bamberg
Bamberg, Germany

Luca Longo
Technological University Dublin
Dublin, Ireland

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-032-08326-5 ISBN 978-3-032-08327-2 (eBook)
https://doi.org/10.1007/978-3-032-08327-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

When Can You Trust Your Explanations? A Robustness Analysis on Feature Importances

Ilaria Vascotto[1(✉)], Alex Rodriguez[1,2], Alessandro Bonaita[3], and Luca Bortolussi[1]

[1] Department of Mathematics, Informatics and Geosciences, University of Trieste, Trieste, Italy
ilaria.vascotto@phd.units.it,
{alejandro.rodriguezgarcia,lbortolussi}@units.it
[2] The Abdus Salam International Center for Theoretical Physics, Trieste, Italy
[3] Assicurazioni Generali Spa, Milan, Italy
alessandro.bonaita@generali.com

Abstract. Recent legislative regulations have underlined the need for accountable and transparent artificial intelligence systems and have contributed to a growing interest in the Explainable Artificial Intelligence (XAI) field. Nonetheless, the lack of standardized criteria to validate explanation methodologies remains a major obstacle to developing trustworthy systems. We address a crucial yet often overlooked aspect of XAI, the robustness of explanations, which plays a central role in ensuring trust in both the system and the provided explanation. To this end, we propose a novel approach to analyse the robustness of neural network explanations to non-adversarial perturbations, leveraging the manifold hypothesis to produce new perturbed datapoints that resemble the observed data distribution. We additionally present an ensemble method to aggregate various explanations, showing how merging explanations can be beneficial for both understanding the model's decision and evaluating the robustness. The aim of our work is to provide practitioners with a framework for evaluating the trustworthiness of model explanations. Experimental results on feature importances derived from neural networks applied to tabular datasets highlight the importance of robust explanations in practical applications.

Keywords: XAI · Robustness · Feature Importance · Neural Networks · Tabular Data · Trustworthy AI

1 Introduction

The popularity of neural networks and their application to high-risk scenarios has recently raised questions on their accountability and trustworthiness. The rapid expansion of the field of Artificial Intelligence (AI) has stimulated legislative

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2578, pp. 225–249, 2026.
https://doi.org/10.1007/978-3-032-08327-2_11

discussions and a series of novel regulations and guidelines has been proposed. In the European Union, the *AI Act* [10] and parts of the *General Data Protection Regulation* (GDPR) [9] have stressed the need for a fairer and more transparent approach to artificial intelligence. Similarly, the United States of America have proposed the new *Blueprint for an AI bill of rights* [31] that strives for a fair development and deployment of AI systems.

Modern AI systems are increasingly complex due to the elevated number of parameters involved to solve challenging tasks and are often referred to as *black boxes* given the opaque nature of their predictions. Transparency is, instead, a fundamental property that AI systems should guarantee, aiming to provide detailed descriptions of the model reasoning, even in natural language. Imagine that a model is being used in healthcare for patient diagnosis. If doctors can understand how it reached a given prediction, they gain a valuable tool for evaluating the correctness of such diagnosis and grow confidence in the AI system, even if they don't fully comprehend the technical aspects of the *black box* model.

The field of Explainable Artificial Intelligence (XAI) has proposed a variety of approaches to *open the black box* and provide explanations addressing the model inner reasoning, for example in the form of feature importances. While not explicitly referred to from a legislative standpoint, XAI can act as a powerful tool in enhancing transparency of AI-based systems. Understanding the reasoning behind a model decision is an asset from a technical standpoint, allowing experts to validate the predictions and detect possible biases before a model is deployed. Additionally, end-users may benefit from explanations as the *right to explanation* cited in the GDPR explicitly requires the user to receive an explanation when the decision is entirely subject to an automated decision system.

Despite their usefulness, the lack of standardized criteria to validate explainability approaches is still a major obstacle towards transparent and trustworthy systems. Although it is a critical aspect, the robustness of explanations remains an often underexplored facet of the development of explanation approaches. Robustness can be defined as an explainer's ability to provide consistent explanations for similar inputs. It can be evaluated through both non-adversarial perturbations, showing intrinsic weaknesses even under small changes, and adversarial attacks, implying a malicious nature of manipulating explanations.

Another complex characteristic of the XAI field is the disagreement problem [14], which occurs in scenarios where multiple explanation methods applied to the same datapoint return contrasting results. The debate over which explanation to choose (or *trust*) is still open, as explanation disagreement posits practical impediments to trustable AI systems.

Our contribution explores the following points:

- We propose a set of desirable properties that a robustness estimator should satisfy and show that our proposal, tailored for feature importance methods, satisfies them all.
- We address the disagreement problem on neural network explanations by proposing an ensemble of explanations focused on the ranking of the features.

- We introduce a framework to test explanation robustness to non-adversarial perturbations and assess their trustworthiness in practical applications.
- We propose a novel validation assessment of the robustness estimation, to tackle the lack of a ground truth.

We have tested our proposal on eight publicly-available tabular datasets and three neural network-specific feature importance methods. Our analysis demonstrates the need for evaluation tools to assess explanation robustness, supporting transparency and accountability in real-world applications.

2 Related Work

LIME [23] and SHAP [16] are among the most widely used XAI techniques in real-world applications. LIME works by fitting an inherently transparent model (such as a linear model or a decision tree) around the datapoint which is being explained. A neighbourhood is constructed from a fixed data distribution, generating a set of points on which the model, which acts as a local explanation, is fitted. SHAP makes use of Shapley values to explain predictions, measuring how the prediction changes when a feature is included or excluded in the feature set. It then averages these changes across all possible combinations of features, producing a vector of feature importances.

Despite their wide use, both methods lack *robustness* (or *stability*), which in this context represents the ability of an explanation method to produce similar and consistent explanations when different conditions change. LIME is an unstable method by design, as the neighbourhood generation step yields different sets of datapoints at each call of the method. This implies that, at each time the method is applied to the same datapoint, a different model is fitted, resulting in explanations whose coefficients differ feature-wise by magnitude or even by sign. SHAP is instead susceptible to feature correlations, sampling variability and data distribution shifts due to the way Shapley values are approximated. The instability of these approaches was first proved in [28], where the authors showed how an adversarial model could easily be defined to mask biased classifiers though unbiased explanations. Their untrustworthiness, along with other model-agnostic additive methods, has also been investigated in [12]. Their theoretical assumptions on feature independence are hardly met in practice, rendering them unable to correctly detect feature interactions when present.

According to [19], explanation robustness can be tested along three directions: robustness to input perturbations, to model changes and to hyperparameter selection. The first one includes the scenarios where the input may be modified by random perturbations or by adversarial attacks to the explanations themselves. The second one refers to manipulations of the model, such as fine-tuning of the parameters with a modified loss function, and the last one considers the influence of technique-specific hyperparameters.

An adversarial attack in the context of XAI is a perturbation of the input such that the model prediction is unchanged but the explanation marks different features as important or not [6]. An evaluation is provided in [11], in

which attacks to interpretations provided by neural networks aim at maximising the change in the explanations (according to different change functions) under the constraint that the perturbation is small and the prediction is unchanged. [8] explores how model manipulation, in particular with the introduction of a penalty term in the loss function, can damage explanations. They call for a rigorous test of robustness to limit the effects of such manipulations. In [4] the authors note that it is possible, for any classifier g, to construct a classifier $\hat{g}$ that exhibits the same behaviour on seen data but presents biased explanations. They propose a version of the presented gradient-based explanations ([5,30]) robust to model manipulation by projecting the explanations on the tangent space of the data manifold.

While random perturbations and adversarial attacks to explanations in the context of images (as in [11]) can be easily examined by human experts, as the changes are often evident even to the naked eye, it is not sufficient to rely on non-quantitative evaluations. Robustness analysis must be supported by an adequate robustness estimation and agreed-upon metrics must be defined. To this end, [13] proposes a robustness score to evaluate explanation robustness when the data generation process is known, but this assumption makes it difficult to adapt to real-world datasets, where a ground truth is hardly available. In [2] the authors propose a formalization of local robustness based on the estimate of the local Lipschitz continuity. They test their proposal on explanations applied to images and show that gradient-based approaches are much more robust than their perturbation counterparts (LIME and SHAP).

The work of [20] presents an in-depth survey on the evaluation of XAI techniques and identifies different metrics that can be used to assess explanation robustness. Similarities can be computed with metrics such as: rank order correlations, top-k intersections, rule matching and structural similarity indexes. [25] proposes similar metrics for evaluating stability, as the Jaccard similarity, additionally requiring that stability tests should be performed using perturbations that do not change the class label and that introduce small amounts of resampling noise to ensure the stability of the explanations.

Robust-by-construction approaches have emerged as an interesting area of research. For example, [3] proposes self-explaining neural networks, a class of models for which faithfulness and robustness are enforced by construction through a specific regularization, aided by a generalization of the above-mentioned local Lipschitz continuity. Similarly, ROPE [15] is a framework based on adversarial training that generates explanations which are robust to both changes in the input and in the data distribution.

We aim at investigating the robustness of explanations to non adversarial perturbations on tabular datasets, presenting a robustness metric that addresses limitations of the previously mentioned methods (recalled within brackets in this paragraph). In particular we define a metric that can be computed even in absence of a ground truth [13], considering practitioners needs (for example, deploying a model without requiring a retraining [8]) and keeping in mind the limitations derived from theoretical assumptions [12]. Importantly, our metric

is bounded within the $[0, 1]$ range, allowing for comparability among methods, datasets and used models [2].

Alongside robustness analysis, we seek to explore the efficacy of an ensemble approach on explanations. [14] presents an overview of the disagreement problem according to practitioners: while it is underlined that the problem is of non-negligible size, the authors do not propose solutions or good practices. In the context of adversarial attacks, [24] showcases the efficacy of ensemble methods as defences on explanations. They test their aggregated explanation on image data and show that it is more resilient to attacks, when a composing explanation method or the model itself is being fooled. Aggregations can also be performed leveraging multiple explanations from the same method: [7] derives a more robust Shapley-value explanation by aggregating the explanations computed on a carefully crafted neighbourhood, minimizing explanation sensitivity. We will consider these characteristics when devising our ensemble and considering the robustness estimator.

3 Background

This section introduces key terminology related to the XAI field and describes the techniques used in the experimental analysis.

3.1 Terminology

While there is not an agreed-upon taxonomy to classify XAI techniques, we follow the proposal of [1], which identifies the following axes of interest:

- Scope of the explanation: a local approach aims at explaining how a given individual prediction is made while a global one focuses on the model as a whole, analysing its overall reasoning.
- Model of interest: model-specific techniques are tailored to the structure of the model under investigation, while model-agnostic ones can be applied to any model.
- Transparency: intrinsically transparent models are interpretable by construction (and are also known as *glass boxes*) while post-hoc techniques are applied after the model is fully trained.

As we will be discussing results obtained from neural networks, we can further distinguish between perturbation-based and gradient-based approaches [26]. Perturbation-based approaches are often model-agnostic and rely either on neighbourhood generation or combinatorial aspects, as in LIME [23] or SHAP [16]. Gradient-based methods, instead, are specific to neural networks and harness their inner structure, mainly taking advantage of the backpropagation mechanisms [5,27,30].

The broader category of approaches we will be considering is that of feature attributions. Having an input $\mathbf{x} = (x_1, \dots x_m)$ with m features, a feature attribution is a vector $\mathbf{a} = (a_1, \dots, a_m)$ of size m where each entry represents

the importance of the corresponding feature towards the model's prediction. According to the specific datatype of $\mathbf{x}$, the attributions may refer to individual variables (tabular data), words or bag of words (natural language), pixels or superpixels (images - in this case explanations are often called *heatmaps*). A positive (negative) sign usually represents a positive (negative) contribution of the feature and its relevance in the prediction. In the following, we will use the terms *feature importances* and *attributions* interchangeably.

3.2 Considered Techniques

We focus on local post-hoc approaches specific to neural networks and that are applicable to nets trained on tabular datasets. We have not considered model-agnostic approaches such as LIME [23] and SHAP [16] due to their known instability [2,12,28], as presented in Sect. 2. Model-specific approaches, such as Saliency and Input X Gradient [26] are also known for being unstable [2,11] and have therefore been excluded from the analysis.

Note that neural network explanations require the selection of the output neuron to be explained: this is more relevant in classification problems where one may want to investigate the features that contributed to any of the classes probability scores. In the following, the target neuron will be chosen as the one associated with the model's predictions, that is, the one with the largest output score.

DeepLIFT. DeepLearning Important Features (DeepLIFT) [27] computes feature importances with respect to their difference from a given reference. In particular, the differences between the two outputs are explained in terms of the differences among the two inputs. Each neuron is analysed with respect to the difference between its activation and that of the reference input. DeepLIFT makes use of contribution scores and multipliers to backpropagate the difference in output through the network. It requires a single forward-backward pass through the net, making it efficient. The propagations are computed through appropriate chain rules, defined according to the neuron's type and its activations.

Integrated Gradients. Integrated Gradients (IG) [30] satisfies the axioms of sensitivity and implementation invariance. It computes the integral of the gradients of a net f along the straight-line path from a baseline x' to the input x, considering a series of linearly separated instances along the path from the baseline to the point of interest. In practice, it takes advantage of an approximation of s steps such that, for the j-th dimension, it holds:

$$\mathrm{IG}_j^{approx} = \frac{x_j - x'_j}{s} \sum_{k=1}^{s} \frac{\partial f(x' + (x - x') \cdot k/s)}{\partial x_j} \tag{1}$$

The authors of [30] found that $s \in (20, 300)$ produced satisfactory approximations but the computation can nonetheless be expensive when the number of steps is large.

Layerwise Relevance Propagation Layerwise Relevance Propagation (LRP) [5] is based on the backpropagation principle. It defines a series of rules to propagate the output score (or *relevance*) $f(x)$ through the net's layers, according to the architecture at hand. The conservation property holds: with R_j the relevance for neuron j, for each pair of layers $\sum_j R_j = \sum_k R_k$ and globally it holds that summing over all layers $\sum_i R_i = f(x)$. Two common propagation rules are the `epsilon` and the `gamma` rules:

$$\begin{aligned} \text{LRP-}\epsilon : R_j &= \sum_k \frac{a_j \cdot w_{jk}}{\epsilon + \sum_{0,j} a_j \cdot w_{jk}} R_k \\ \text{LRP-}\gamma : R_j &= \sum_k \frac{a_j \cdot (w_{jk} + \gamma \cdot w_{jk}^+)}{\epsilon + \sum_{0,j} a_j \cdot (w_{jk} + \gamma \cdot w_{jk}^+)} R_k \end{aligned} \tag{2}$$

where a_j is the activation of neuron j, w_{jk} is the weight linking neuron j to neuron k in the following layer (w_{jk}^+ is a positive weight), $\sum_{0,j}$ is the sum over all lower-layer activations.

Missingness Property. DeepLIFT, Integrated Gradients and LRP satisfy the missingness property: if $x_j = 0 \Rightarrow a_j = 0$. The features corresponding to the null entries in the feature vector will have null coefficients in the attribution vector, representing a lack of importance towards the prediction. This property is particularly relevant when dealing with categorical variables, preprocessed with one-hot encoding.

4 Methodology

Our approach is applied to tabular datasets and classification problems. Let us introduce the following notation: let $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ be a dataset with N datapoints and m features such that $(\mathbf{x}^i, y^i) = (x^{(i,1)}, \ldots, x^{(i,j)}, \ldots, x^{(i,m)}, y^i)$ with y^i a class label. The dataset is split into a training, validation and test datasets, identified by $\mathcal{D}_{train}$, $\mathcal{D}_{valid}$ and $\mathcal{D}_{test}$ respectively. Let $f(\cdot)$ be a neural network trained on $\mathcal{D}_{train}$ and t be a target class. Let e be an explanation method (or explainer) and $e(\mathbf{x}^i) \coloneqq e(\mathbf{x}^i, f)$ the explanation of model f prediction of point $\mathbf{x}^i$. More specifically, let the feature attribution vector of the l-th method be $\mathbf{a}_l^i = (a_l^{(i,1)}, \ldots, a_l^{(i,m)})$.

4.1 Robustness Estimator

We define the robustness of an explanation as a measure of its variability when the input is modified. If we consider $\mathbf{x}$ the original datapoint, $\tilde{\mathbf{x}}$ a perturbation, e an explanation method and $e(\mathbf{x})$ the corresponding explanation, then:

$$\mathbf{x} \to \tilde{\mathbf{x}}, e(\mathbf{x}) \to e(\tilde{\mathbf{x}}) \Rightarrow r(\mathbf{x}, e) = g(\mathbf{x}, \tilde{\mathbf{x}}, e) \tag{3}$$

that is, the robustness r of $e(\mathbf{x})$ is a function of the chosen explanation method, the original datapoint and the perturbed one.

Further considering a constraint of the form $dist(\mathbf{x}, \tilde{\mathbf{x}}) < \epsilon$ with $\epsilon > 0$ and *dist* a distance metric (such as the euclidean distance), we can introduce the notion of local robustness.

Definition 1. *An explanation method e is locally robust if perturbing the input results in a similar explanation. If* $\mathbf{x} \rightarrow \tilde{\mathbf{x}}$ *with* $dist(\mathbf{x}, \tilde{\mathbf{x}}) < \epsilon$ $(\epsilon > 0) \Rightarrow e(\mathbf{x}) \approx e(\tilde{\mathbf{x}})$.

Given a robustness estimator $\hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i) \coloneqq \hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i, e, f)$ that measures the robustness of the explanation method e applied to the neighbourhood $\mathcal{N}^i$ of the point $\mathbf{x}^i$ over the model f, we can define the following set of *desiderata*.

Property 1. If $r(\mathbf{x}^i, \tilde{\mathbf{x}}^i) \coloneqq r(\mathbf{x}^i, \tilde{\mathbf{x}}^i, e, f)$ is the robustness of $e(\mathbf{x}^i)$ with respect to the perturbation $\tilde{\mathbf{x}}^i$, then the robustness $\mathcal{R} = \mathbb{E}[r]$ is estimated by:

$$\hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i) = \frac{1}{|\mathcal{N}^i|} \sum_{\tilde{\mathbf{x}}^i \in \mathcal{N}^i} r(\mathbf{x}^i, \tilde{\mathbf{x}}^i) \tag{4}$$

where $\mathcal{N}^i = \{\tilde{\mathbf{x}}^i | \tilde{\mathbf{x}}^i = \mathbf{x}^i + \lambda \text{ with } dist(\mathbf{x}^i, \tilde{\mathbf{x}}^i) < \epsilon \ (\epsilon > 0),\ \lambda \in \mathbb{R}^m\}$.

Two points which are close to each other within the dataspace will produce explanations with comparable robustness scores.

Property 2. If $\hat{\mathcal{R}}$ is a local robustness estimator, then for two distinct points $\mathbf{x}^i, \mathbf{x}^j$ such that $dist(\mathbf{x}^i, \mathbf{x}^j) < \epsilon$ $(\epsilon > 0)$ it holds that:

$$\exists\, \delta > 0 \text{ s.t. } |\hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i) - \hat{\mathcal{R}}(\mathbf{x}^j, \mathcal{N}^j)| < \delta \tag{5}$$

Robustness estimation is intrinsically linked to uncertainty in the estimates, due to errors in the explanations themselves and the lack of a ground truth for the robustness.

Property 3. $\hat{\mathcal{R}}$ is such that $\hat{\mathcal{R}} = \mathcal{R} + \theta_\epsilon$ where $\mathcal{R} = \mathbb{E}[r]$ is the true robustness and $\theta_\epsilon \neq 0$ is an error term.

By definition, the neighbourhood generation is highly influential in the robustness estimation process when non-adversarial perturbations are considered. On-manifold perturbations better reflect the true data distribution and the manifold learned by the model, therefore exhibit greater robustness scores than random perturbations which may be off-manifold.

Property 4. If $\mathcal{N}^i$ is an on-manifold neighbourhood and $\bar{\mathcal{N}}^i$ an off-manifold one, then $\hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i) > \hat{\mathcal{R}}(\mathbf{x}^i, \bar{\mathcal{N}}^i)$.

The robustness of an aggregation of explainers is bounded by the robustness of the individual components.

Property 5. If the explainer $\bar{e}$ is an aggregation of explainers, $\bar{e} = agg(e_1, \ldots, e_l)$, then, if $\hat{\mathcal{R}}(e) \coloneqq \hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i, e, f)$, it holds that $\hat{\mathcal{R}}(\bar{e}) \leq \max(\hat{\mathcal{R}}(e_1), \ldots, \hat{\mathcal{R}}(e_l))$.

Two equivalent models predicting the same class for a datapoint will be related to explanations with comparable robustness scores. The data manifold learned by the models influences the robustness score.

Property 6. Let $f(\cdot)$ and $g(\cdot)$ be two models with comparable accuracy on the dataset $\mathcal{D}$. If the point $\mathbf{x}^i$ is predicted to belong to the same class by both models, say $\hat{y}_f^i = \hat{y}_g^i$, and let $\hat{\mathcal{R}}(f) \coloneqq \hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i, e, f)$, then:

$$|\hat{\mathcal{R}}(f) - \hat{\mathcal{R}}(g)| < \delta \text{ with } \delta > 0 \tag{6}$$

We propose the robustness r to be computed as $r(\mathbf{x}^i, \tilde{\mathbf{x}}^i) = \rho(e(\mathbf{x}^i, f), e(\tilde{\mathbf{x}}^i, f))$, where ρ is the Spearman's rho rank correlation coefficient and $e(\mathbf{x}^i, f)$ the explanation of model f prediction of point $\mathbf{x}^i$. By Property 1, it then holds that the robustness $\mathcal{R}$ can be estimated via:

$$\hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i, e, f) = \frac{1}{|\mathcal{N}^i|} \sum_{\tilde{\mathbf{x}}^i \in \mathcal{N}^i} \rho(e(\mathbf{x}^i, f), e(\tilde{\mathbf{x}}^i, f)) \tag{7}$$

where $\mathcal{N}^i = \{\tilde{\mathbf{x}}^i | \tilde{\mathbf{x}}^i = \mathbf{x}^i + \lambda \text{ with } \lambda \in \mathbb{R}^m,\ dist(\mathbf{x}^i,\ \tilde{\mathbf{x}}^i) < \epsilon\, (\epsilon > 0) \text{ and } \hat{y}^i = \hat{y}^{\tilde{i}}\}$. By definition, it holds that $0 \leq \hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i, e, f) \leq 1$.

We will show in Sect. 5 that our estimator also satisfies Properties 2-6.

4.2 Neighbourhood Generation

As remarked in Property 4, neighbourhood generation is an influential step in the estimation of the robustness, as the score is averaged over the set $\mathcal{N}^i$. In the previous subsection, we have limited the constraints of the neighbourhood to only consider perturbed points $\tilde{\mathbf{x}}^i$ which are close to the original datapoint $\mathbf{x}^i$ and for which the model's prediction is the same, $\hat{y}^i = \hat{y}^{\tilde{i}}$. We will consider two possible neighbourhood generation mechanisms which deeply influence the robustness computation. Let us distinguish between numerical and categorical variables, x_{num}^i and x_{cat}^i respectively, as they require different perturbations by construction.

Random Neighbourhood ($\mathcal{N}_R$). A first naive approach is the random generation of the neighbourhood, consisting of the addition of random white noise to numerical variables and a random flip of the categorical ones:

$$\begin{cases} \tilde{x}_{num}^i = x_{num}^i + \delta^i \text{ with } \delta^i \leftarrow \mathcal{N}(0, \sigma^2) \\ \tilde{x}_{cat}^i = \text{flip}(x_{cat}^i) \text{ with probability } \gamma_{cat} \end{cases} \tag{8}$$

The flip of a categorical variable entails a random sampling among the possible modalities associated with that variable, the observed value of x_{cat}^i being excluded.

Medoid-Based Neighbourhood ($\mathcal{N}_M$). We propose a more refined mechanism that leverages the manifold hypothesis to generate perturbed datapoints

which are still on-manifold, as our interest lies in testing non-adversarial perturbations consistent with the observed data distribution. Consider a k-medoid clustering on $\mathcal{D}_{valid}$, with $k_{medoids}$ selected so that each cluster is, on average, of size $n_k = 10$. For each cluster, the medoid $\mathbf{x}^c$ is used as a representative and its k_M nearest neighbours among the other cluster centres are stored in the set $NN^c = (\mathbf{x}^1, \ldots, \mathbf{x}^{k_M})$. For each point $\mathbf{x}^i \in \mathcal{D}_{test}$ we want to test, the associated cluster c is retrieved. From the corresponding cluster centre neighbours list NN^c, one of the medoids is randomly chosen, say $\mathbf{x}^M$. With α and α_{cat} the probabilities of perturbing a numerical and a categorical variable respectively, a perturbation is performed according to the following scheme:

$$\begin{cases} \tilde{x}^i_{num} = (1 - \bar{\alpha}) \cdot x^i_{num} + \bar{\alpha} \cdot x^M_{num} \text{ with } \bar{\alpha} \leftarrow Beta(\alpha \cdot 100, (1 - \alpha) \cdot 100) \\ \tilde{x}^i_{cat} = \begin{cases} x^i_{cat} \text{ with probability } 1 - \alpha_{cat} \\ x^M_{cat} \text{ with probability } \alpha_{cat} \end{cases} \end{cases} \tag{9}$$

With both generating schemes, the resulting neighbourhood should be of at least size $n = 100$ to ensure statistical significance. A filtering step is then performed to remove the perturbations for which the model prediction is different from $f(\mathbf{x}^i)$. Hyperparameter tuning is performed on $\theta_1 = (\sigma, \gamma_{cat})$ and $\theta_2 = (\alpha, \alpha_{cat}, k_M)$ to ensure that, on average, at least 95% of the points are kept.

The main differences among the two approaches are presented in Fig. 1, where the Swiss roll dataset is used as an example. Both schemes were applied to the same datapoint: the left most column represents a 3D visual of the dataset (in the shape of a rolled piece of paper), while the middle one is a view from above. It is easy to note that the random neighbourhood is expanded beyond the Swiss roll spiral shape. This is more evident in the right-most column, where a zoomed-in visualization is proposed: while the random neighbourhood does not follow the data manifold even when a small perturbation is applied ($\sigma = 0.05$), the medoid-based one remains constantly within bounds despite a larger coefficient being used ($\alpha = 0.3$).

4.3 Ensemble

We present a novel aggregation approach that aims at dealing with the disagreement problem [14] by merging feature importances, focusing on the ranking of the features according to their absolute value. Let L be the number of methods which are being aggregated and let $\mathbf{r}^i_l$ be the l-th ranking, a vector of size m storing the indices that would sort the array $|\mathbf{a}^i_l|$ in decreasing order. Then, let an average attribution $\mathbf{a}^i_{ens}$ be defined by:

$$a^{(i,j)}_{ens} = \frac{\sum_{l=1}^{L} r^{(i,j)}_l \cdot w^{(i,j)}_l}{\sum_{l=1}^{L} w^{(i,j)}_l} \cdot (1 + \lambda \bar{n}^{(i,j)}) \tag{10}$$

where w_l is the weight corresponding to the l-th feature attribution vector, $\lambda = 0.15$ is a penalization term and $\bar{n}^{(i,j)}$ is the number of methods for which there is

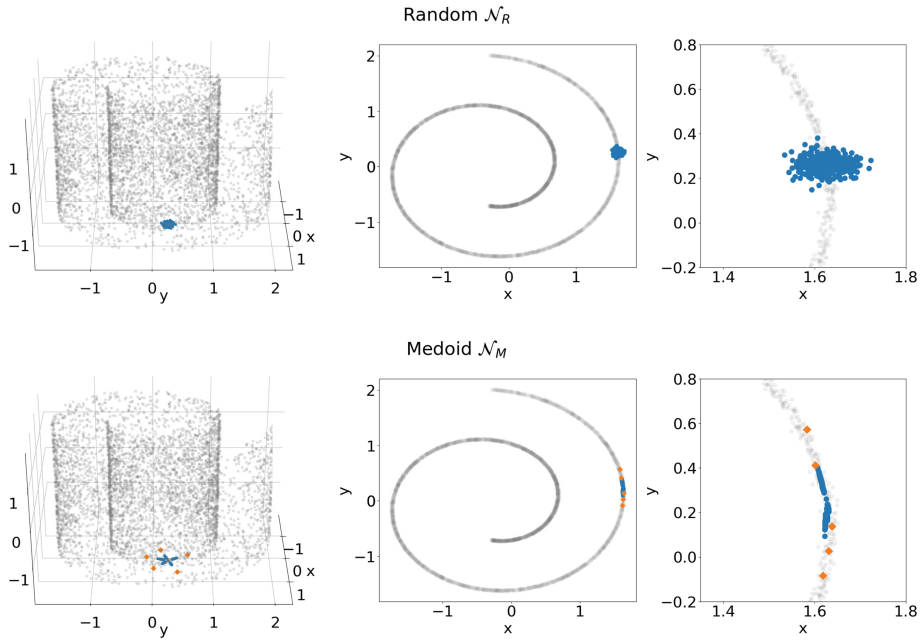

Fig. 1. An example of neighbourhood generation ($n = 500$) on the Swiss roll dataset: the top row depicts the random neighbourhood generation $\mathcal{N}_R$ ($\sigma = 0.05$) while the bottom one the medoid-based generation $\mathcal{N}_M$ ($\alpha = 0.3$ and $k_M = 5$). Blue datapoints represent the generated perturbations while the orange ones in the bottom row are the k_M neighbouring cluster centres.

a disagreement on the sign of the attribution $a_l^{(i,j)}$ for $l = 1, \ldots, L$. The ranking $\mathbf{r}_{ens}^i$ storing the indices that would sort the attribution vector $\mathbf{a}_{ens}^i$ in increasing order is the results of our approach.

This aggregation method is able to deal with practical issues that emerge when considering multiple explanations altogether. First of all, attributions may result in coefficients very close, but not equal, to zero. A zero coefficient implies that the feature is not important towards the prediction, but the lack of a common scale for feature attributions makes it difficult to discriminate between important and unimportant features only based on the absolute value of the corresponding coefficient. We aim at limiting this issue by considering the following weighting scheme where, for the l-th methods, i-th point and j-th feature, it holds:

$$w_l^{(i,j)} = \frac{\sigma(\mathbf{a}_l^i)}{\sqrt{|\overline{\mathbf{a}_l^i}| \cdot |a_l^{(i,j)}|}} \tag{11}$$

with $\mathbf{a}_l^i$ the feature importance vector, $\overline{\mathbf{a}_l^i}$ its average and $\sigma(\mathbf{a}_l^i)$ the standard deviation.

Smaller values of the feature importance coefficients translate into larger weights, that contribute to moving a feature into the set of non-informative ones of our ranking, effectively capturing the scarcer importance towards the prediction. As coefficients may differ greatly in magnitude between multiple methods, a normalization is then introduced with the consideration of the average and standard deviation within the weighting scheme.

The ensemble is computed as the weighted average of the $L = 3$ rankings and a penalization term λ is introduced to penalize features that exhibit sign disagreement among the L methods, as we consider it a symptom of feature instability. The penalization term allows us to favour the features for which there is sign concordance ($\bar{n}^{(i,j)} = 0$), either positive or negative, and when the attribution magnitude is non negligible. Considering how larger weights move features towards the set of non influential ones, the penalization contributes to a lower attribution value for concordant features, effectively capturing their (absolute) relevance and importance towards the prediction.

We will compare our ensemble to another aggregation method in Sect. 5. Inspired by [24], we will simply consider the average of the attributions, when the vectors have norm 1. If $\mathbf{a}_l^i$ is such that $\|\mathbf{a}_l^i\|_2 = 1$, then

$$a_{mean}^{(i,j)} = \frac{1}{L} \sum_{l=1}^{L} a_l^{(i,j)} \tag{12}$$

with $\mathbf{a}_{mean}^i$ such that $\|\mathbf{a}_{mean}^i\|_2 = 1$.

One of the disadvantages of using the mean as aggregation is that it may assign a zero attribution even when all L coefficients are non-null. Assume that, for the j-th feature, the $L = 3$ coefficients derived from the corresponding methods are $(v, v, -2v)$ with $v > 0$. The mean will be equal to zero, implying that the feature is non relevant towards the prediction, but the importance for each of the L method highlights a relevance of magnitude at least v. Our ensemble is, instead, able to take this into account, penalizing the disagreement but still considering the feature to have some level of relevance towards the prediction.

While we have computed both aggregations with the feature attributions derived from DeepLIFT, Integrated Gradients and LRP (as presented in Subsect. 3.2), both approaches can be easily extended to include a higher number of feature attribution methods.

4.4 When Can You Trust Your Explanations?

Assessing the trustworthiness of explanations on previously unseen datapoints is a non trivial task. While it is possible to compute an estimate of the robustness via Eq. 7, we argue that the result may not reflect the true robustness of the considered datapoint, as it may lay in an unstable area of the feature space. We want to verify not only if a datapoint is robust, but also if it lies in a robust area of the feature space: knowing that the neighbourhood is non robust allows us to *doubt* the robustness score of a previously-unseen datapoint. To tackle

this issue, we propose the usage of a k-nearest neighbours regressor fitted on $\mathcal{D}_{valid}$ robustness scores. For $\mathbf{x}^i \in \mathcal{D}_{test}$, we compute both the robustness score $\hat{\mathcal{R}}(\mathbf{x}^i) \coloneqq \hat{\mathcal{R}}(\mathbf{x}^i, \mathcal{N}^i, e, f)$ via 7 and the regressor's prediction $\mathcal{R}_{knn}(\mathbf{x}^i)$. According to a selected threshold value r_{th}, it is possible to discriminate between three scenarios:

1. if $\hat{\mathcal{R}}(\mathbf{x}^i) \geq r_{th}$ and $\mathcal{R}_{knn}(\mathbf{x}^i) \geq r_{th}$ then $\mathbf{x}^i$ is a robust point;
2. if $\hat{\mathcal{R}}(\mathbf{x}^i) \geq r_{th}$ and $\mathcal{R}_{knn}(\mathbf{x}^i) < r_{th}$ then $\mathbf{x}^i$ is an uncertain point, as it lies in an uncertain area of the feature space and its robustness should be carefully considered;
3. if $\hat{\mathcal{R}}(\mathbf{x}^i) < r_{th}$ then $\mathbf{x}^i$ is a non robust point.

The second scenario represents a set of conditions that practitioners should carefully evaluate: despite the robustness score being greater than the selected threshold, the local information derived from the neighbours suggests otherwise. This scenario aims at ringing a bell in the practitioner evaluating a given explanation: knowing that it may be misleading, the analysis will be more careful and require a more detailed human-evaluation of both the prediction and its explanation.

The usage of a $k$nn regressor instead of a $k$nn classifier allows for independence from the selected robustness threshold r_{th}, requiring the model be fitted only once. The number of neighbours k_R is a dataset-dependent hyperparameter and relies on the goodness of approximation of the robustness score by the regressor.

The selection of the threshold r_{th} is a delicate step of the procedure: as the robustness estimator is bounded by construction in the $[0, 1]$ range, it is possible to select a case-specific threshold to discriminate between robust and non robust datapoints. As it will be seen in Subsect. 5.1, in which hyperparameter selection is discussed in detail, a default value of $r_{th} = 0.80$ works well on most of the datasets.

4.5 A Complete Pipeline

With the considerations presented up to this point, we can discuss the complete framework (Fig. 2) for robustness evaluation and explanation trustworthiness.

1. Split the dataset into $\mathcal{D}_{train}, \mathcal{D}_{valid}, \mathcal{D}_{test}$ and perform the required preprocessing steps.
2. Train a neural network on $\mathcal{D}_{train}$.
3. Perform k-medoid clustering on $\mathcal{D}_{valid}$ and compute the k_M nearest neighbours among the medoids.
4. For each point $\mathbf{x}^j \in \mathcal{D}_{valid}$, generate a neighbourhood $\mathcal{N}^j$ of size n.
5. Compute the attributions of DeepLIFT, Integrated Gradients and LRP and merge them following the ensemble aggregation scheme.
6. Compute the robustness score via Eq. 7 for each point $\mathbf{x}^j \in \mathcal{D}_{valid}$.
7. Use the previously computed robustness scores to train a k-nearest neighbours regressor and select an appropriate threshold r_{th}.

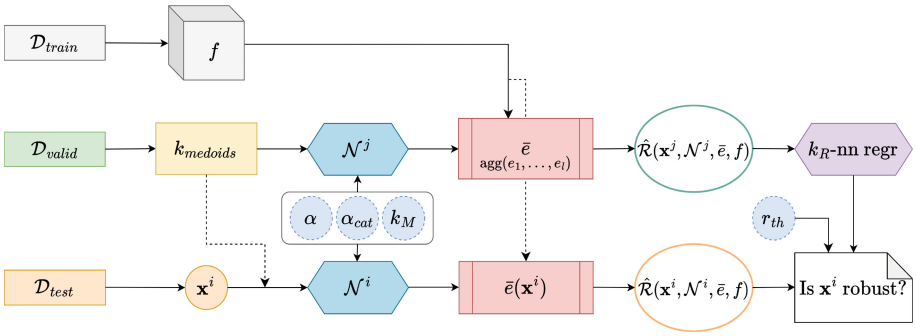

Fig. 2. The complete pipeline.

8. For each datapoint $\mathbf{x}^i \in \mathcal{D}_{test}$, predict the medoid cluster it belongs to and generate a neighbourhood $\mathcal{N}^i$.
9. Compute the L feature attribution vectors and the resulting aggregation.
10. Compute then the robustness score via Eq. 7 and the predicted robustness via the $k$nn regressor.
11. Assess the trustworthiness following Subsect. 4.4.

Note that the computational cost of the procedure is dominated by the computation of the individual attributions, as a pass through the network is required for each datapoint, for each perturbation and for each of the three methods. The scaling of the underlying robustness method inherits this complexity.

The advantages of using this framework include the leveraging of an ensemble-based explanation, which captures the signals of L individual explanation methods, and the estimation of the robustness on a carefully constructed neighbourhood. The pipeline is designed to point out possibly non robust points, even when they appear so, ensuring greater trustworthiness in the system and a more aware analysis from a practitioners perspective.

The presented framework is applicable to datasets entirely made of numerical variables, but to correctly consider also categorical ones a small adjustment is required. Categorical features are often preprocessed with a one-hot encoding before a neural network is trained. For example, a feature x_{cat} with four modalities is represented by a vector of the form $(0, 0, 1, 0)$, where the non zero-entry corresponds to the observed modality. A feature attribution method applied to the one-hot encoded feature vector returns a set of importances of the form $(0, 0, u, 0)$: the non-zero entry is the only one associated with a non-null value. This is to be expected as the attribution methods we are evaluating satisfy the missingness property (Subsect. 3.2). To limit the effects of zero-entries in the Spearman's rho computation, we propose a *reverse encoding* of categorical variables. They are represented by the observed modality and the variable x_{cat} is associated with an attribution $a_l = u$. This allows us to consider feature vectors of size m, as in the original feature vector, and to perform a more effective com-

parison of the robustness. The *reverse encoding* is applied in steps 5 and 9, prior to the ensemble's computation.

4.6 Validation

Robustness estimation is often subject to the lack of a ground truth, as the data generation process is not known in real-world applications and comparison against expert-provided explanations can be unfeasible due to the large amount of data being examined. We argue that robust (explanation-wise) points lie in a robust area of the feature space and can be deemed robust even when passed though different models. Our assumption implies that non robust points will exhibit differences both in the explanations and in the predictions of multiple models, as their lack of robustness is a somewhat intrinsic characteristic of the area of the manifold in which they lie.

Let us consider three neural networks, say f_1, f_2 and f_3, which have comparable accuracy over $\mathcal{D}_{train}$ and that differ either in the number of hidden layers or neurons per layer. Let $\mathcal{D}_{agree}$ be the subset of datapoints for which models f_1, f_2, f_3 predict the same class and $\mathcal{D}_{disagree}$ the subset for which one of the models predicts a different class. Consider a point to be robust if $\hat{\mathcal{R}}(\mathbf{x}^i) \geq r_{th}$ and non-robust otherwise.[1] We propose the validation to follow a ROC/AUC analysis, where the True Positive Rate (TPR) and False Positive Rate (FPR) are defined as:

$$\text{TPR} = \frac{\#\{\text{Robust \& Agree}\}}{\#\{\text{Agree}\}} \qquad \text{FPR} = \frac{\#\{\text{Robust \& Disagree}\}}{\#\{\text{Disagree}\}} \tag{13}$$

Varying the threshold value r_{th}, it is possible to plot the ROC curve and compute the corresponding AUC value.

5 Experimental Evaluation

5.1 Experimental Setting

We have selected the following publicly-available datasets from the UC Irvine Machine Learning Repository: `beans`, `cancer`, `mushroom`, `white wine`, `adult` and `bank marketing`. We have additionally used the `heloc` and `ocean` datasets, following the work of [21,29] respectively. The first four represent toy examples, as the classification tasks are easier to tackle even with non-neural models and present an overall lower number of both features and datapoints. All the datasets propose binary or multiclass classification tasks and contain both numerical and categorical variables.

We have relied on the Python libraries `pytorch` and `captum` for the implementation of our approach. The former was used for the training and usage of

[1] Note that, in this case, we are not considering the classification presented in Subsect. 4.4, but only if the robustness score $\hat{\mathcal{R}}$ is above the selected threshold r_{th}.

the nets while the latter, a `pytorch`-compatible explainability framework developed by Meta, was applied for the retrieval of the attribution vectors. The full implementation is available on Github.[2]

For all datasets, the following preprocessing steps were performed:

- Standardization of numerical variables and one-hot encoding of the categorical ones.
- Removal of correlated features. Spearman's rank correlation coefficient was used for numerical variables while the normalized mutual information criterion for the categorical ones. The removal of highly correlated features ensures that neighbourhood generation is more aligned with the data distribution and that the computed attributions are non distorted by unconsidered correlations among the features.
- Softmax in the final layer of the net, also for binary classification examples. It ensures stability during attribution computations, as higher relevances are backpropagated. This also reduces the effects of the vanishing gradient problem, when the attribution is returned as a zero-vector, as the signal is lost when the output score is propagated through the net.
- Selection of the `gamma` rule for LRP attributions. It was chosen as it was the rule minimizing the vanishing gradient problem, and it was applied to all the layers of the nets (as they are all linear layers).

As anticipated in Subsect. 4.6, we trained three neural networks per dataset, say Model 1, 2 and 3, with comparable accuracy scores (Appendix A, Table 4). Model 1 represents the baseline model, model 2 has more layers and more neurons in each layer while model 3 is a more compact version, with fewer layers and neurons. The ReLU activation function was used in all cases, with the exception of the `ocean` dataset, where tanh was used, following the structure in [29].

Hyperparameter Selection. Default values for the neighbourhood generation hyperparameters are $\theta_1 = (\sigma = 0.05, \gamma_{cat} = 0.05)$ for the random generation $\mathcal{N}_R$ and $\theta_2 = (\alpha = 0.05, \alpha_{cat} = 0.05, k_M = 5)$ for the medoid-based one $\mathcal{N}_M$. In both cases, the neighbourhood should be of size at least $n = 100$ (we have set $n = 100$ in our experiments) and dataset-specific hyperparameters are set via a grid search, ensuring that at least 95% of the generated datapoints are kept within the neighbourhood, i.e. they are predicted to belong to the same class as the original datapoint.

The number of neighbours k_R to be used in the $k$nn regressor is chosen as the one minimizing the approximation error over the robustness scores derived from all three nets in each dataset. A good default value is $k_R = 7$.

The robustness threshold r_{th} is selected by looking at the distribution of robust, non robust and uncertain datapoints at varying levels of the threshold. In particular, it is chosen as the threshold value that corresponds to the first inflection point of the robust percentage curve. The default value $r_{th} = 0.80$ works well in most scenarios.

[2] https://github.com/ilariavascotto/XAI_robustness_analysis

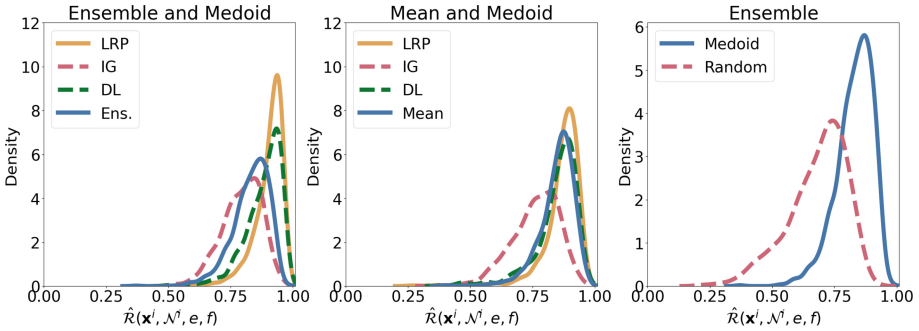

Fig. 3. The robustness score distribution for `adult` dataset of the ensemble (left) and of the mean (centre) with medoid-based neighbourhood generation; random and medoid-based neighbourhood generations are compared for the ensemble aggregation (right).

Dataset-specific hyperparameters can be found in Appendix A, Table 5.

5.2 Results

We will begin the discussion focusing on the results derived from Model 1, that acts as our baseline, on the test set $\mathcal{D}_{test}$.

Figure 3 depicts the robustness score distribution derived from the ensemble (left) and the mean (centre) aggregations with the medoid-based neighbourhood generation mechanism on the `adult` dataset. They are compared with the robustness scores derived from the individual XAI approaches, taking into account the nature of the aggregation. In particular, the ensemble is compared to the robustness computed on the feature importance vectors in absolute value, to mimic the reasoning of the ensemble's construction, while the mean is compared to the feature vectors with sign. As stated by Property 5, in both cases the aggregated explanation robustness acts as an average of the robustness of the individual approaches and is limited by their span. In this example, Integrated Gradients is on average the least robust method while DeepLIFT and LRP present grater values of the estimated robustness. Note that this behaviour is dataset dependent: Integrated Gradients is not, in general, the least robust method. The aggregated explanations, either with the ensemble or the mean, represent an advantage over the use of individual approaches. In particular, the aggregation acts as a conservative explanation, that takes into account the individual method's robustness and their feature-wise agreement. While the average robustness may be lower than that of some of the considered approaches, it is able to take into account the possible undesirable effects of a less robust and disagreeing method, flagging possible untrustworthiness for a given datapoint.

The rightmost part of Fig. 3 depicts the comparison between the effects of two neighbourhood generating schemes on the ensemble non-adversarial robustness. As previously shown in Fig. 1, the random neighbourhood consists of datapoints which are off-manifold, while the medoid-based one is constructed to remain on-

Table 1. When can you trust your explanations? Comparison of the robustness scores with ensemble and mean aggregations with $r_{th} = 0.80$.

		Ensemble			Mean		
Dataset	$\mathcal{D}_{test}$	Robust	Uncertain	Non Robust	Robust	Uncertain	Non Robust
beans	500	74.6%	7.2%	18.2%	93.0%	2.2%	4.8%
cancer	50	62.0%	12.0%	26.0%	100.0%	0.0%	0.0%
mushroom	400	0.0%	1.2%	98.8%	0.0%	0.0%	100.0%
white wine	200	7.5%	26.5%	66.0%	56.5%	11.0%	32.5%
adult	1000	63.4%	7.1%	29.5%	72.6%	7.3%	20.1%
bank	1000	65.2%	7.1%	27.7%	44.6%	16.4%	39.0%
heloc	500	62.8%	9.2%	28.0%	60.6%	14.0%	25.4%
ocean	10000	79.1%	5.6%	15.3%	53.5%	16.0%	30.5%

manifold, mimicking the effects of non-adversarial perturbations. This reflects into larger average robustness over the test set of the latter neighbourhood over the random one, validating that the robustness estimator with the ensemble consistently satisfies Property 4. The same property is satisfied also by the mean aggregation and the individual XAI methods when tested individually.

Table 1 presents the classification of the test set into robust, uncertain and non robust datapoints (as per Subsect. 4.4). For comparability, we set $r_{th} = 0.80$ for all datasets and both aggregation methods, even if the ensemble often requires lower values of r_{th} compared to the mean (see Appendix A). Medoid-based neighbourhoods were used, as it was shown in Fig. 3 that they produce larger robustness scores.

For almost all the datasets, robust datapoints are the majority, with the exception of the `white wine` dataset with the ensemble and `mushroom` dataset with both aggregations. This is due to the selection of the threshold value equal for all datasets, as carefully selected dataset-specific thresholds would be lower than 0.80. The percentages of uncertain and non robust datapoints are non-negligible. In particular, we consider uncertain datapoints to be the most interesting ones to investigate. They represent areas of the feature space where the robustness estimation is uncertain (as per Property 3) and should be carefully considered during a practical evaluation.

In Fig. 4 we present a two dimensional visualization of the UMAP [18] projections of the validation set $\mathcal{D}_{valid}$, clustered with HDBSCAN algorithm [17], where each cluster is coloured by its mean robustness score. UMAP is a dimensionality reduction technique that allows us to visualize a lower dimensional projection of the data. We applied the density-based clustering algorithm HDBSCAN to such projections, producing clusters without the need of setting hyperparameters for the desired number of clusters, as in k-means for example. As can be seen in the figure, the projected data space may be more or less complex according to the dataset being analysed. Robustness homogeneity within the derived clusters

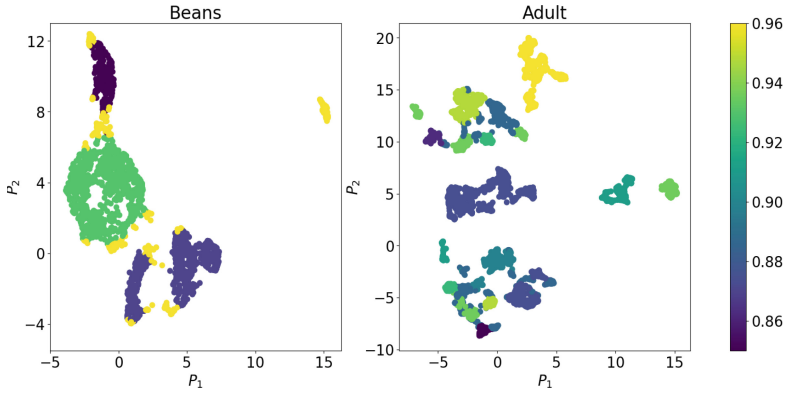

Fig. 4. HDBSCAN clusters, coloured by mean robustness scores, over the two dimensional UMAP projections (P_1, P_2): `beans` dataset (left) and `adult` dataset (right).

Table 2. Comparison of ensemble's robustness classification (with $r_{th} = 0.80$) according to Model 1,2,3 concordance in the prediction.

	Robust			Non Robust		
Dataset	Agree	Disagree	N. points	Agree	Disagree	N. points
beans	95.35%	4.65%	409	95.60%	4.40%	91
cancer	91.89%	8.11%	37	92.31%	7.69%	13
mushroom	100.00%	0.00%	5	100.00%	0.00%	395
white wine	94.11%	5.88%	68	92.42%	7.58%	132
adult	93.76%	6.24%	705	80.68%	19.32%	295
bank	95.99%	4.01%	723	90.25%	9.75%	277
heloc	82.22%	17.78%	360	56.43%	43.57%	140
ocean	85.21%	14.8%	8475	80.33%	19.67%	1525

support the claims of Property 2, in which close points are expected to exhibit smaller differences between their robustness scores. This allows us to support the use of the $k$nn regressor to estimate local robustness scores as points are naturally grouped in clusters with similar robustness scores.

As introduced in Subsect. 4.6, the validation of robustness estimations is subject to the lack of a ground truth. Table 2 shows how the percentage of points over which the three models (Model 1, 2 and 3) (dis)agree varies according to the predicted robustness of the datapoints, with $r_{th} = 0.80$ for all datasets. The ensemble robustness is computed with the medoid-based neighbourhood. Note that the `mushroom` dataset is a particular case, as all three methods reach 100% accuracy and are, therefore, always agreeing. It can be seen that the percentage of disagreeing points within the non-robust ones is, for most datasets, greater than that of the robust ones. This supports our validation proposal, as we consider

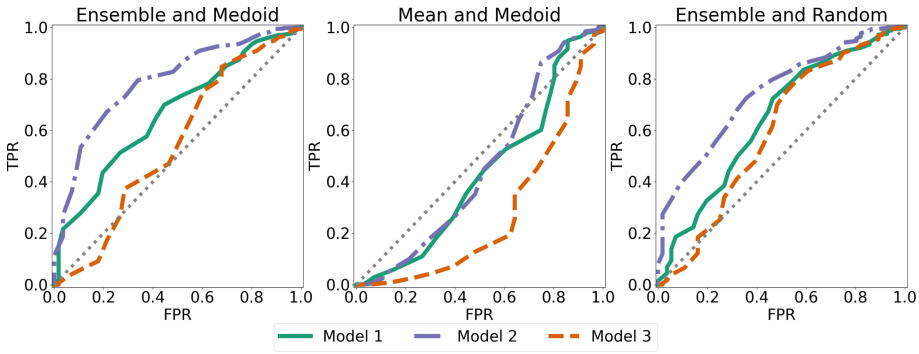

Fig. 5. ROC curves on the `bank` dataset with ensemble aggregation and medoid-based neighbourhood (left), mean aggregation and medoid-based neighbourhood (centre) and ensemble aggregation with random neighbourhood (right). The gray dotted line represents the bisecting line.

Table 3. Average AUC value of Model 1, 2, 3 for both aggregation types and neighbourhoods. For each dataset, the largest average AUC is presented in bold.

	Medoid		Random	
Dataset	Ensemble	Mean	Ensemble	Mean
beans	**0.5711**	0.4854	0.4842	0.4067
cancer	0.7346	**0.7400**	0.6766	0.6612
mushroom	0.0000	0.0000	0.0000	0.0000
white wine	0.4762	0.5931	0.4819	**0.6614**
adult	0.7018	0.6695	**0.8284**	0.8077
bank	**0.6670**	0.3883	0.6612	0.4696
heloc	0.6640	**0.6673**	0.6262	0.5875
ocean	**0.5194**	0.5128	0.4350	0.4053

the disagreement in the predictions to be a symptom of non robustness, as per Property 6.

The ROC/AUC analysis presented in Subsect. 4.6 allows us to jointly consider the aggregation method and the neighbourhood generation scheme which better fit the dataset at hand. Figure 5 depicts the ROC curves of the three models in varying scenarios: Model 2 is consistently associated with the highest ROC curve, suggesting that, for the `bank` dataset, a deeper net is better able to propose robust explanations. Moreover, the ROC curves are preferable for all three models with the ensemble aggregation and the medoid-based neighbourhood, while the mean aggregation (as shown in the central plot) presents ROC curves even below the bisecting line.

This is further confirmed by the results presented in Table 3, where the average AUC is computed for the four possible combinations of aggregation method and neighbourhood generation. The dataset-wise maximum values of AUC are presented in bold, showing how the ensemble is, on average, preferable with respect to the mean aggregation and how the medoid-based neighbourhood is related to larger AUC values in most examples. The `mushroom` dataset represents, as in Table 2, a peculiar case: having only agreeing datapoints the ROC and AUC cannot be computed and are therefore exempt from this analysis.

6 Conclusions and Future Work

We presented a novel framework to test explanation robustness, introducing a new neighbourhood generation mechanism, an ensemble approach to merging explanations and a validation test. We have shown that our robustness estimator satisfies the *desiderata* 1-6 and it overcomes the limitations of other metrics (Sect. 2). In practical applications, our approach would aid practitioners in understanding the quality of an explanation in terms of its robustness, allowing questioning on the proposed results when the framework flags a datapoint as *uncertain*.

We have proposed our work targeting neural networks, but the approach is agnostic in nature with respect to both the investigated model and the XAI techniques being applied. Future steps include a first generalization of the proposal on different classes of machine learning models, such as tree-based ones, assuming that local feature importance approaches are available for testing. Future work will also aim at better investigating the relationship between robustness and adversarial attacks. In particular, we aim at assessing the defence ability of our ensemble - as aggregated explanations have proved to be more resilient to adversarial attacks in numerous contexts - and to validate whether our robustness estimator is able to detect attacks. Lastly, we wish to investigate how our explanations could be used to increase the robustness of classifiers, as in [22].

Acknowledgments. This study was partially carried out within the PNRR research activities of the consortium iNEST funded by the European Union Next-GenerationEU (PNRR, Missione 4 Componente 2, Investimento 1.5âĂŞ D.D. 1058 23/06/2022, ECS 00000043).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Dataset Description and Hyperparameters

Table 4. Dataset description and model accuracy.

	Dataset details							Accuracy (%) - $\mathcal{D}_{train}$		
Dataset	Classes	#Num	#Cat	$\mathcal{D}_{train}$	$\mathcal{D}_{valid}$	$\mathcal{D}_{test}$	$k_{medoids}$	Model 1	Model 2	Model 3
beans	7	7	0	10888	2223	500	225	93.41	93.44	89.47
cancer	2	15	0	397	121	50	10	99.5	99.75	99.24
mushroom	2	0	21	6498	1225	400	120	100.00	100.00	100.00
white wine	2	9	0	3918	780	200	80	89.23	89.10	86.55
adult	2	5	7	36177	8045	1000	1000	91.39	91.38	91.09
bank	2	5	9	36168	8043	1000	1000	91.99	91.76	91.45
heloc	2	14	2	8367	1592	500	130	85.50	85.57	85.01
ocean	6	8	0	109259	30328	10000	3000	92.04	87.88	92.27

Table 5. Dataset-specific hyperparameter selection.

	$\mathcal{N}_M$							$\mathcal{N}_R$					
	Neighbourhood			Ensemble		Mean		Neighbourhood		Ensemble		Mean	
Dataset	α	α_{cat}	k_M	k_R	r_{th}	k_R	r_{th}	σ	γ_{cat}	k_R	r_{th}	k_R	$r_{th}-$
beans	0.10	–	10	9	0.85	9	0.90	0.02	–	5	0.75	7	0.80
cancer	0.10	–	4	11	0.85	9	0.90	0.10	–	5	0.55	7	0.65
mushroom	–	0.15	10	7	0.70	11	0.60	–	0.15	9	0.70	11	0.60
white wine	0.15	–	5	11	0.85	7	0.80	0.03	–	9	0.45	7	0.65
adult	0.05	0.05	5	9	0.80	9	0.80	0.05	0.05	5	0.70	5	0.70
bank	0.05	0.10	5	7	0.80	11	0.80	0.05	0.10	7	0.75	9	0.75
heloc	0.05	0.05	5	15	0.80	13	0.80	0.03	0.10	11	0.45	5	0.60
ocean	0.05	–	5	5	0.65	5	0.75	0.001	–	5	0.75	5	0.65

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
2. Alvarez-Melis, D., Jaakkola, T.: On the robustness of interpretability methods (June 2018). https://doi.org/10.48550/arXiv.1806.08049
3. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining neural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS 2018, pp. 7786–7795. Curran Associates Inc., Red Hook (2018). https://hdl.handle.net/1721.1/137669

4. Anders, C., Pasliev, P., Dombrowski, A.K., Müller, K.R., Kessel, P.: Fairwashing explanations with off-manifold detergent. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, 13–18 Jul, vol. 119, pp. 314–323. PMLR (2020), https://proceedings.mlr.press/v119/anders20a.html
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10(7)** (2015). https://doi.org/10.1371/journal.pone.0130140
6. Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: a survey. Inform. Fusion **107**, 102303 (2024). https://doi.org/10.1016/j.inffus.2024.102303
7. Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020 (2021). https://dl.acm.org/doi/abs/10.5555/3491440.3491857
8. Dimanov, B., Bhatt, U., Jamnik, M., Weller, A.: You shouldn't trust me: learning models which conceal unfairness from multiple explanation methods. In: ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 2473–2480. IOS Press (2020). https://doi.org/10.3233/FAIA200380
9. European Commission: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj
10. European Commission: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts (COM(2021) 206 final) (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
11. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(01), pp. 3681–3688 (2019). https://doi.org/10.1609/aaai.v33i01.33013681
12. Gosiewska, A., Biecek, P.: Do not trust additive explanations. arXiv: Learning (2019). https://doi.org/10.48550/arXiv.1903.11420
13. Kantz, B., et al.: Robustness of explainable artificial intelligence in industrial process modelling. In: ICML 2024 Workshop ML for Life and Material Science: From Theory to Industry Applications (2024). https://doi.org/10.48550/arXiv.2407.09127
14. Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. Trans. Mach. Learn. Res. (2024). https://doi.org/10.21203/rs.3.rs-2963888/v1
15. Lakkaraju, H., Arsov, N., Bastani, O.: Robust and stable black box explanations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, JMLR.org (2020). https://proceedings.mlr.press/v119/lakkaraju20a/lakkaraju20a.pdf

16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Adv. Neural Inform. Process. Sys. **2017**, 4766—-4775 (2017). https://dl.acm.org/doi/10.5555/3295222.3295230
17. Malzer, C., Baum, M.: A hybrid approach to hierarchical density-based cluster selection. In: 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 223–228 (2020). https://doi.org/10.1109/MFI49285.2020.9235263
18. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: uniform manifold approximation and projection. J. Open Source Softw. **3**(29), 861 (2018). https://doi.org/10.21105/joss.00861
19. Mishra, S., Dutta, S., Long, J., Magazzeni, D.: A survey on the robustness of feature importance and counterfactual explanations (2023). https://arxiv.org/abs/2111.00358
20. Nauta, M., et al.: From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable ai. ACM Comput. Surv. **55**(13s), 1–42 (2023). https://doi.org/10.1145/3583558
21. Pawelczyk, M., Broelemann, K., Kasneci, G.: Learning model-agnostic counterfactual explanations for tabular data. In: Proceedings of The Web Conference, WWW 2020, p. 3126–3132. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3366423.3380087
22. Rasouli, P., Yu, I.C.: Analyzing and improving the robustness of tabular classifiers using counterfactual explanations. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1286–1293 (2021). https://doi.org/10.1109/ICMLA52953.2021.00209
23. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" explaining the predictions of any classifier. In: NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session, pp. 97—-101 (2016). https://doi.org/10.18653/v1/n16-3020
24. Rieger, L., Hansen, L.K.: A simple defense against adversarial attacks on heatmap explanations (2020). https://arxiv.org/abs/2007.06381
25. Rosenfeld, A.: Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS. vol. 1, pp. 45–50 (2021). https://dl.acm.org/doi/abs/10.5555/3463952.3463962
26. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: a review of methods and applications. Proc. IEEE **109**(3), 247–278 (2021). https://doi.org/10.1109/JPROC.2021.3060483
27. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: 34th International Conference on Machine Learning, ICML 2017, vol. 7, pp. 4844–4866 (2017). https://dl.acm.org/doi/10.5555/3305890.3306006
28. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: adversarial attacks on post hoc explanation methods. In: AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, p p. 180–186 (2020). https://doi.org/10.1145/3375627.3375830
29. Sonnewald, M., Lguensat, R.: Revealing the impact of global heating on north atlantic circulation using transparent machine learning. J. Adv. Model. Earth Syst. **13** (2021). https://doi.org/10.1029/2021MS002496

30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: 34th International Conference on Machine Learning, ICML 2017 **7**, 5109–5118 (2017). https://dl.acm.org/doi/10.5555/3305890.3306024
31. The White House: Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People (2022). https://www.whitehouse.gov/ostp/ai-bill-of-rights/