ABDUCTIVE EXPLANATIONS FOR GROUPS OF SIMILAR SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Explaining the decisions of machine learning models is crucial as their use becomes widespread. While many approaches to explanation are based on heuristics or surrogate models without formal guarantees, formal explanations provide reasoning for a particular decision that is guaranteed to be valid. We focus on abductive explanations (AXp) that identify sufficient subsets of input features for a given classification. We extend AXp to not only cover a particular sample, but to cover all of the samples whose features are within a given interval, providing explanations that remain valid even when the features in the explanation vary by up to δ. In addition to applying this notion of *δ-robust AXp* to a single sample, we also consider *group explanations* (δ-gAXp), which give a common explanation for a group of samples that share the same classification. We evaluate our approach by producing explanations for neural networks with the help of Marabou, a neural network verifier. The evaluation shows that, compared to a recent approach for finding a maximally "inflated" explanation, a δ-robust AXp covers a significant volume of the inflated explanation with a dramatically lower runtime. Our evaluation also provides evidence that group explanations capture important features for all the samples within the group much faster than computing explanations for each sample separately.

1 INTRODUCTION

Many real world applications are increasingly relying on decision-making systems based on machine learning, including in sensitive contexts such as health care, law enforcement and finance (Karimi et al., 2023). This has spurred calls for a right to explanation for those affected by the decisions of automated decision-making systems. For example, OECD AI Principles recommend that AI systems should "provide information that enable those adversely affected by an AI system to challenge its output"[1] while the data protection regulation of the European Union GDPR requires that users subject to automatic decision-making have the right to "meaningful information about the logic involved"[2]. Indeed, research into methods for explaining machine learning models has blossomed in recent years (Confalonieri et al., 2021; Minh et al., 2022; Bodria et al., 2023; Marques-Silva, 2024).

Many proposed explanation methods, however, can produce misleading explanations (Kumar et al., 2020; Huang & Marques-Silva, 2023a; Marques-Silva & Huang, 2024). Approaches such as the popular SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) are heuristic and thus do not provide formal guarantees. *Formal* explanations, in contrast, produce verifiably valid explanations (Shih et al., 2018; Ignatiev et al., 2019; Marques-Silva & Ignatiev, 2022). In this paper, we focus specifically on formal *abductive* explanations (AXps) (Marques-Silva & Ignatiev, 2022), which explain why the model classifies a sample as it does. More concretely, an abductive explanation is a subset of the features of an input sample such that it is guaranteed that the classifier assigns the same label to any sample sharing the values of these features with the input sample. AXps, in other words, offer a local feature selection explanation.

Abductive explanations suffer from a notable limitation, in that they are fragile to small perturbations in feature values. An AXp is only valid for the exact feature values of a given sample, so after a

[1] https://oecd.ai/en/dashboards/ai-principles/P7
[2] https://gdpr-text.com/read/article-13/

minimal perturbation to a single feature it may no longer be valid, even though the perturbed sample is almost identical to the original. This locality can be a limiting factor in real-world applications where measurements are noisy or uncertain. For instance, in medical diagnosis, an explanation for a classification based on a patient's blood pressure being 140.0 mmHg might be invalid for a reading of 140.1 mmHg, although the difference is a slight measurement error. Such a brittle explanation is intuitively inferior to an explanation that holds for all similar values. Furthermore, one might be interested in explanations for a *group* of similar samples. For example, given five patients with similar health records and a similar classification given by the model, it might be more instructive to see explanations for the group than for the individuals.

In this work we consider the extension of AXps to a neighbourhood of samples rather than to a single sample. We introduce δ-robust AXps, where δ is a robustness bound that ensures that an abductive explanation remains valid when the feature values vary within δ. A separate δ_i can be set for each feature i. With these robustness bounds, we lift AXps from a local to semi-local explanations where explanations are valid for a group of samples. We also introduce a type of group abductive explanations δ-gAXps, which explain groups of similar samples by providing a δ-robust AXps for a representative of the group. When applied to a single sample, our approach is similar to a recent approach to compute *inflated* abductive explanations (Izza et al., 2024b; 2025), which give a maximal interval for each feature within which an AXp is valid. In contrast, our approach first sets a desired interval and then finds a valid AXp for this potentially smaller set of intervals, hugely reducing the computational cost. Given the difficulty of producing formal explanations, it is crucial to find the most efficient methods for obtaining the desired results.

We make the following contributions. We introduce δ-robust AXps, an extension of abductive explanations that provide formal guarantees while being robust to small perturbations, and δ-gAXp, a formalization of abductive explanations for groups of samples. We implement our approach using a state-of-the-art neural network verifier Marabou (Wu et al., 2024), leveraging the formal guarantees and computational efficiency of constraint-solving. We show in a comprehensive empirical evaluation that our approach scales effectively, adding little overhead compared to computing standard AXps. We show that, compared to (maximally) inflated AXps, δ-robust AXps cover a significant portion of the feature space at a fraction of the computational cost. Finally, we evaluate δ-gAXps based on their induced formal feature attribution scores (FFA) (Yu et al., 2023) and show that they effectively capture the feature importance within a group. Using the FFA scores of each individual sample of the group as the ground truth, δ-gAXps perform better than SHAP and LIME, and comparable to a brute-force method of computing the average of the FFAs for all samples of the group, with an orders of magnitude lower runtime.

2 Related Work

Explainable AI has gained significant attention in recent years (Confalonieri et al., 2021; Minh et al., 2022; Bodria et al., 2023), leading to the development of diverse methodologies including heuristic explanations, surrogate models, feature attribution, feature selection, and interpretable models. In this work, we focus on formal explanations, especially feature selection.

Researchers have approached formal explanations from various perspectives. Among these, feature selection methods, which identify subsets of features that constitute explanations, can be roughly divided into *abductive* and *counterfactual* explanations (Wachter et al., 2018; Shih et al., 2018; Ignatiev et al., 2019; Izza & Marques-Silva, 2021; Ignatiev & Marques-Silva, 2021; Parmentier & Vidal, 2021; Wu et al., 2023b; 2024; Karimi et al., 2020; Audemard et al., 2023; Marques-Silva & Ignatiev, 2022; Marques-Silva, 2024; Karimi et al., 2023; Guidotti, 2024). Intuitively, abductive explanations answer the question *why this classification* and counterfactual explanations answer *how to obtain a different classification*. Formally, an abductive explanation is a subset of a sample's features such that any sample fixing these features to their respective values receives the same classification as the original sample. A counterfactual explanation, conversely, is a subset of features such that any sample with different values for these features receives a different classification than the original sample.

Robust explanations have been studied for counterfactual explanations (Jiang et al., 2023; 2024), but they have received less attention for AXps. For AXps, a type of robustness called distance-restricted AXps (Izza et al., 2024a; Huang & Marques-Silva, 2023b) and ϵ-robust explanations (Wu

et al., 2023a) has been considered, where the *relevant* features, i.e. features that are part of the explanation, must remain fixed while the *irrelevant* features, i.e. features not in the explanation, can only vary within a given bound ϵ.

In contrast to distance-restricted explanations, our δ-robust AXps extend AXps by remaining valid when the relevant features vary by up to δ, regardless of how much the irrelevant features vary. In the parlance of Jiang et al. (2024), our δ-robust AXps are robust against noisy execution. As mentioned above, the closest to this sense of robustness for AXps are inflated AXps (Izza et al., 2024b), which give bounds to the relevant features so that explanations remain valid when the input variations occur inside the bounds. The algorithm to generate inflated AXps maximizes the bounds of each feature in a given AXp, one feature at a time, with multiple calls to a solver that checks whether the current explanation is valid. This approach has two drawbacks compared to our δ-robust AXps. First, the feature(s) whose bounds are maximized last can be left with tiny bounds or no bounds at all. Second, they are computationally costly since they require multiple calls to a validity checker.

Group explanations present a separate but related challenge, since they attempt to explain multiple samples with the same classification simultaneously. They have recently gained attention for counterfactual explanations (Warren et al., 2023; Carrizosa et al., 2024a;b; Wielopolski et al., 2024), but, to the best of our knowledge, group explanations have not been considered for AXps. Our work addresses these gaps in the literature by considering *robust abductive* explanations, and applying them to produce abductive explanations for *groups of samples*.

3 Abductive Explanations

Let $f : \mathbb{R}^d \to \{1, \ldots, k\}$ be a classifier that maps samples $\mathbf{x} = \{x_1, \ldots, x_d\} \in \mathbb{R}^d$ to classes $c \in \{1, \ldots, k\}$. An abductive explanation (AXp) answers the question *why does f classify $\mathbf{x}$ as c* by identifying a subset of features E, such that the classifier assigns the label c to any sample that shares all values of E with $\mathbf{x}$. Formally, a subset of features $E \subseteq \mathbb{F}$ is an AXp of $\mathbf{x}$ if

$$\forall \mathbf{z} \in \mathbb{R}^d : \bigwedge_{i \in E} (z_i = x_i) \implies f(\mathbf{z}) = f(\mathbf{x})$$

and there is no proper subset of E satisfying this condition.

3.1 δ-robust Abductive Explanations

A shortcoming of an abductive explanation (AXp) (Ignatiev et al., 2019) is that a small change to one of the relevant features (i.e. a feature in the explanation) can make the explanation invalid. In many real world scenarios, measurement errors can occur, so the fragility of formal explanations limits their practicality. To address this limitation, we propose a variant of abductive explanations that is valid when relevant features vary within specified bounds. This new formulation not only makes abductive explanations robust to input variation, but it allows us to define formal explanations for groups of samples, clarifying the behaviour of classifiers on a group of samples rather than just one.

A δ-robust AXp for a sample $\mathbf{x}$ with classification $f(\mathbf{x}) = c$ is a subset of features $E \subseteq \mathbb{F}$ and a set of robustness bounds $\{[x_i - \delta_i, x_i + \delta_i] \mid i \in E\}$, such that $f(\mathbf{z}) = c$ for any input $\mathbf{z} \in \mathbb{R}^d$ where $z_i \in [x_i - \delta_i, x_i + \delta_i]$ for all $i \in E$. The robustness parameter $\boldsymbol{\delta} = \{\delta_i\}_{i \in E}$ controls the maximum allowed variation in each relevant feature. Formally, a pair $(E, \boldsymbol{\delta})$ is a δ-robust AXp if

$$\forall \mathbf{z} \in \mathbb{R}^d : \bigwedge_{i \in E} (z_i \in [x_i - \delta_i, x_i + \delta_i]) \implies f(\mathbf{z}) = f(\mathbf{x})$$

and there is no proper subset of E satisfying this condition.

This definition ensures that abductive explanations remain valid when relevant features vary within their specified bounds, addressing the fragility to input variation of standard abductive explanations. For convenience, when $\delta_i = v$ for each $i \in E$, we may write that $\delta = v$ and refer to a v-robust AXp. Note that δ-robust AXps are a generalization of AXps. If $(E, \boldsymbol{\delta})$ is a δ-robust AXp for a sample $\mathbf{x}$, then E is an AXp for $\mathbf{x}$. Similarly, an AXp is a 0-robust AXp.

Example 1. *Figure 1 illustrates an AXp and a δ-robust AXp with $\delta = 0.09$. In both, the irrelevant features are the same and can take whatever value without changing the classification. In the relevant features, however, the AXp shows a precise point where the explanation is valid, represented*

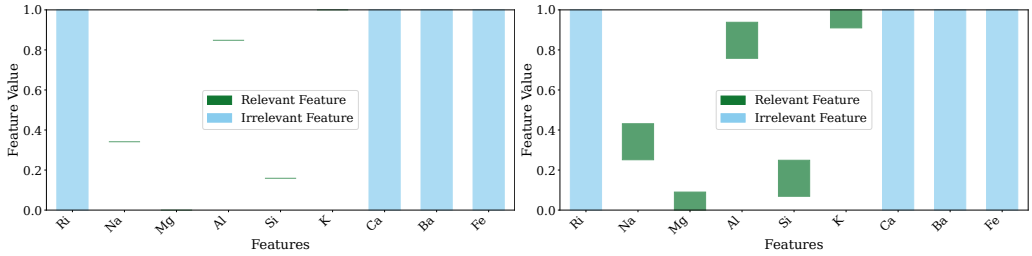


Figure 1: Comparison of two explanations for the first sample with class 5 of the Glass dataset. An AXp (left) and a δ-robust AXp for $\delta = 0.09$ (right).

by the narrow green line corresponding to the feature values of the input sample, while the δ-robust AXp shows intervals within which the relevant features can take any value without the classification changing, represented by the large green rectangles corresponding to the intervals of validity.

To compute a δ-robust AXp, we assume that features are normalized to $[0, 1]$ and that we have a formal model of the classifier in question f as a set of logical constraints and an oracle for checking whether a given set of constraints is satisfiable. In our implementation we use Marabou, a formal verifier of properties in neural networks based on constraint solving (Wu et al., 2024). We impose further constraints as the formula ϕ during the course of the algorithm. The algorithm takes as input a sample $\mathbf{x}$ and a set of robustness bounds $\boldsymbol{\delta} = \{\delta_i\}_{i=1}^{d}$, and finds a set E such that $(E, \{\delta_i\}_{i \in E})$ is a δ-robust AXp for $\mathbf{x}$.

In Algorithm 1, we modify the standard deletion-based algorithm for computing minimal unsatisfiable sets of constraints (Chinneck & Dravnieks, 1991) to compute a δ-robust AXp. First, if the classification is not guaranteed even with all features fixed to be within their respective bounds, there are no δ-robust AXps and the algorithm terminates (lines 2–3). Otherwise, include all features $\mathbb{F}$ are included in an over-approximation of a δ-robust AXp E, and each feature i is considered in turn to determine whether to include it in E (lines 5–9). The order of iteration can be arbitrary, but some orders might result in smaller or otherwise more desirable explanations (Wu et al., 2023a); we discuss this more in the Appendix. For each feature i, the algorithm sets constraints so that i can take any value, and other features j can take values based on whether it is in E or already deemed irrelevant. The features in E are the potentially relevant features, so they must be within their robustness bounds (line 6). An exception is the feature i, which we allow to take any value in order to check whether it is irrelevant (lines 6–7). Features not in E are irrelevant features, so they can take any value in $[0, 1]$ (line 7). In line 8 the function GUARANTEED$(\phi, f(\mathbf{z}) = f(\mathbf{x}))$ uses a constraint-solving oracle and returns `True` iff $f(\mathbf{z}) = f(\mathbf{x})$ holds under all values of $\mathbf{z}$ that satisfy the constraints ϕ. If this is the case, then the feature i can demonstrably take any value without changing the classification, so i is irrelevant and it is removed from E (line 9).

3.1.1 COMPARISON TO INFLATED ABDUCTIVE EXPLANATIONS

Inflated abductive explanations (Izza et al., 2024b) are a related approach to our δ-robust AXps. They maximize the bounds of relevant features one feature at a time and separately for the lower and upper

Algorithm 1 Computing a δ-robust AXp for sample $\mathbf{x}$

1: **function** COMPUTEROBUSTAXP($\mathbf{x}$, $\boldsymbol{\delta}$)
2: **if** $\neg$GUARANTEED$(\{(|z_j - x_j| \leq \delta_j) \mid j \in \mathbb{F}\}, f(\mathbf{z}) = f(\mathbf{x}))$ **then**
3: **return** `False`
4: $E \leftarrow \mathbb{F}$
5: **for** each feature i in $\mathbb{F}$ **do**
6: $\phi \leftarrow \{(|z_j - x_j| \leq \delta_j) \mid j \in E \setminus \{i\}\}$
7: $\phi \leftarrow \phi \cup \{(z_j \in [0, 1]) \mid j \in \mathbb{F} \setminus (E \setminus \{i\})\}$
8: **if** GUARANTEED$(\phi, f(\mathbf{z}) = f(\mathbf{x}))$ **then**
9: $E \leftarrow E \setminus \{i\}$
10: **return** E

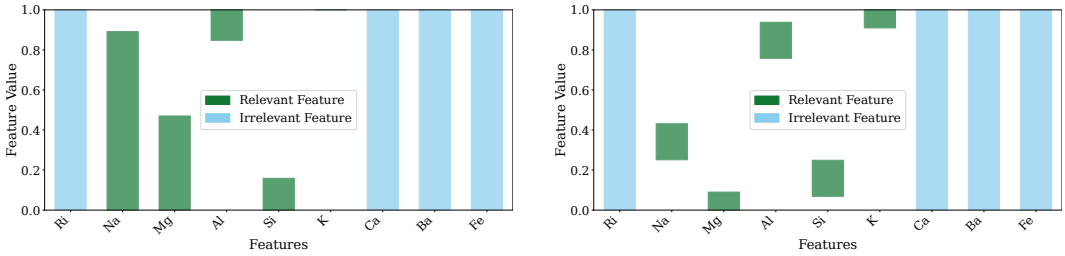


Figure 2: Comparison of two explanations for the first sample with class 5 of the Glass dataset. An inflated AXp (left) generated by inflating a δ-robust AXp for $\delta = 0.09$ (right).

bound. They achieve this by first computing an AXp and then considering each feature i in the AXp in turn. They increase the bound b_i for feature i by a small, predetermined amount η (precision) and check whether the explanation is still valid with a call to an oracle (typically a solver for a constraint language such as propositional satisfiability or mixed-integer linear programming). When the explanation is no longer valid, they take the previous bound (i.e. $b_i - \eta$) and repeat the process for the lower bound of feature i. The result is an explanation together with a set of maximal bounds on each relevant feature that guarantee that varying the values of the relevant feature within these bounds does not change the classification. In other words, inflated AXps are a similar to δ-robust AXps with a separate δ for each feature and each direction (i.e. increasing the value and decreasing the value) where all the δs are maximal. Computing an inflated AXp is more computationally demanding than computing a δ-robust AXp, since to inflate, an oracle needs to be called potentially hundreds of times. Additionally, the features considered last in computing an inflated AXp might get very low bounds.

Example 2. *In Figure 2 we illustrate an inflated AXp and a δ-robust AXp sharing the same underlying AXp. The bounds of the inflated AXp are larger in general, but since inflated AXps are maximized one feature at a time, some features have smaller bounds. Notably, the bounds for feature 'K' could not be inflated at all and instead the feature must remain fixed to a single value in the inflated AXp. Conversely, the δ-robust AXp guarantees given bounds for each feature, so every feature has flexibility. This δ-robust AXp covers 43% of the area of the inflated AXp.*

3.2 Group Explanations

While AXps explain the classification of a single sample, in many cases it is insightful to explain the classification of a group of samples. In medical diagnosis, for example, researchers might want to understand what features are important to diagnose a condition in general, not just for a particular patient. Group explanations address this need by identifying the features that explain why the samples in a group receive the same classification. Group explanations can be thought of as semi-local, since they do not apply to a single sample (local) or to the whole model (global), but to a group of samples. They are most effective with samples that are naturally grouped together, such as patients with similar symptoms, customers with similar profiles, or products with similar characteristics.

In Algorithm 2, we illustrate how to compute a type of group explanation, δ-gAXp, by computing δ-robust AXps on a representative of the group with robustness bounds that depend on the variance of feature values occurring in the group. The algorithm takes a group of samples $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ that share the same classification and a vector $\boldsymbol{\delta}$ as input. First, it computes a representative of the group by averaging the samples in the group (line 2). If the representative receives a different clas-

Algorithm 2 Computing a δ-gAXp for a group of samples with the same classification

1: **function** ComputeGroupAXp($X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\boldsymbol{\delta}$)
2: $\quad\bar{\mathbf{x}} \leftarrow \text{Mean}(X)$
3: $\quad$**if** $f(\bar{\mathbf{x}}) \neq f(\mathbf{x}_1)$ **then return** `False`
4: $\quad\boldsymbol{\gamma} \leftarrow \boldsymbol{\delta} \cdot \text{StandardDeviation}(X)$
5: $\quad E \leftarrow$ ComputeRobustAXp($\mathbf{x}, \boldsymbol{\gamma}$)
6: $\quad$**return** $(E, \boldsymbol{\gamma})$

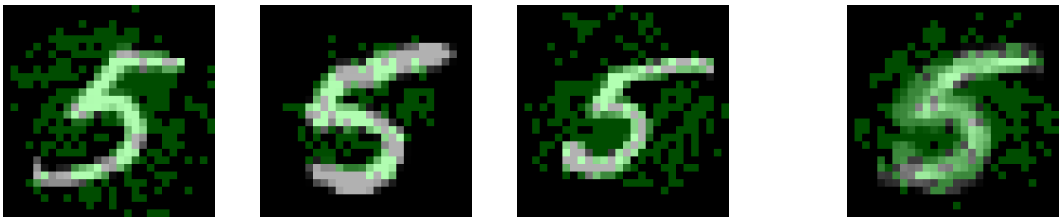


Figure 3: From left to right, the first three images represent individual AXps for three images in MNIST, while the last image represents a δ-gAXp for $\delta = 0.3$ for the former images. The green pixels represent the relevant features in the explanations.

sification than the samples in X, the group is too dissimilar and has no δ-gAXp (line 3). Otherwise, the robustness bounds of δ_i are computed so that they are proportional to the standard deviation of each feature i within the group for the representative (line 4). Finally, it checks whether a δ-robust AXp exists with Algorithm 1 using these bounds (line 5). If the δ-robust AXp exists, we call the output a δ-gAXp $= (E, \boldsymbol{\gamma})$. For any sample $\mathbf{z} = \{z_1, \ldots, z_d\}$ it holds that if $|z_i - \bar{x}_i| \leq \gamma_i$ for every feature $i \in E$, then $f(\mathbf{z}) = f(\bar{\mathbf{x}})$.

The parameter $\boldsymbol{\delta}$ multiplies the standard deviation of the features in the group producing the final bounds $\boldsymbol{\gamma}$ for the features. This formulation sets wider bounds for features that have diverse values in the group and narrower bounds for features that have similar values in the group. If $\delta = 0$, the δ-gAXp is a AXp for the group representative, while larger values of δ increase the robustness bounds, making the explanation valid for a wider range of samples.

Example 3. *In Figure 3, we compute an AXp for three MNIST samples and their common δ-gAXp (for $\delta = 0.3$). The individual explanations show different patterns for each sample, reflecting their specific characteristics, while the δ-gAXp captures the common features that are important across the three samples, reflecting the general characteristics that make them belong to the same class.*

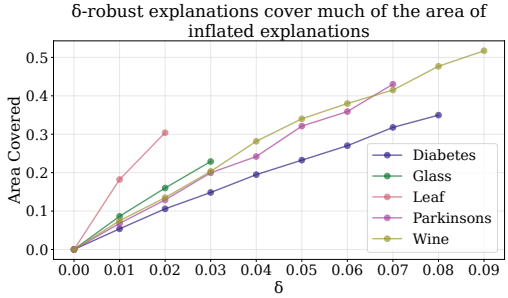
4 Empirical Evaluation

We conducted an empirical evaluation using 2x12 core Xeon E5 2680 v3 2.50GHz CPUs with 128GB DDR4-2133 of RAM. The code for the experiments can be found in the supplementary material and will be made available in open source.

Datasets We evaluate our methods on both *classic machine learning datasets* Glass (OpenML, a), Leaf (OpenML, b), Parkinsons (OpenML, c), Diabetes (Efron et al., 2004) and Wine (Aeberhard & Forina, 1992) and *synthetic datasets* generated using scikit-learn (function `make_classification`). The classic datasets represent real-world scenarios with varying dimensionality and complexity, while the synthetic datasets allow us to control dimensionality and feature characteristics for scalability analysis. In the synthetic datasets, we have 10000 samples, 2 classes, and 10 to 25 features. The `make_classification` function generates datasets with informative and noisy features. Informative features are those that contribute to the target variable, while noisy features are random and do not provide useful information for classification. We set the number of informative features to $\lfloor n/2 \rfloor$, where n is the total number of features.

Machine Learning Model We train a fully-connected neural network (NN) with 2 hidden layers of 10 neurons each and ReLU as their activation function for each dataset. We train each NN for 500 epochs with a learning rate of 0.001, a batch size of 16, and a patience of 10. We use 0 as the seed. We compute explanations with Marabou (Wu et al., 2024), which encodes the NN into a logical formula and allows for formally verifying properties of the NN.

Comparisons We compare δ-robust AXps against inflated AXps (Izza et al., 2024b) (computed with a reimplementation using Marabou) as the primary baseline, since they represent the most relevant alternative approach. We evaluate the total volume of feature values covered by both approaches. For δ-gAXps, i.e. group explanations, we compare the formal feature attribution (FFA) (Yu et al., 2023) induced by our approach against SHAP, LIME, and a brute-force approach of



	Diabetes	Glass	Leaf	Parkinsons	Wine
$\delta = 0.00$	1.1574	1.1088	**1.6919**	**1.2455**	**1.0777**
$\delta = 0.01$	1.1400	**1.0847**	1.7071	1.3002	1.1018
$\delta = 0.02$	1.1345	1.1936	1.7253	1.4251	1.1106
$\delta = 0.03$	**1.1272**	1.1345	-	1.4121	1.1217
$\delta = 0.04$	1.1592	-	-	1.3869	1.1181
$\delta = 0.05$	1.1600	-	-	1.4919	1.1250
$\delta = 0.06$	1.1511	-	-	1.4730	1.1665
$\delta = 0.07$	1.1578	-	-	1.4327	1.1723
$\delta = 0.08$	1.2342	-	-	-	1.1515
$\delta = 0.09$	-	-	-	-	1.1874
Inflated	12.1076	4.5886	14.5489	34.9010	14.9624

Figure 4: Across 100 samples, average area of inflated AXps covered by δ-robust AXps (left) and average runtime (right).

computing explanations separately for each sample in the group and aggregating them afterwards. We consider the FFA scores of the individual samples as the ground-truth. We also evaluate the runtime since all the explanation problems we consider are computationally hard.

4.1 Evaluation of δ-robust Explanations

We evaluate our approach across 100 samples by comparing the volume of a δ-robust AXp to the volume of an inflated AXp based on the same underlying AXp (i.e. sharing the same relevant features), and by comparing their runtimes. We use a precision of $\eta = 0.01$ for the inflated AXp. To present cases where applications could realistically use δ-robust AXps, we only show data for values of δ where at least 50% of the samples have a δ-robust AXp.

Figure 4 shows the results. On the left side we see that δ-robust AXps achieve up to 20–50% of the coverage of inflated AXps, and on the right that they are 4 to 30 times faster to compute than inflated AXps. Thus, even though δ-robust AXps do not maximize the validity bounds of each feature, they provide valid explanations covering a substantial percentage of the area of inflated AXps at a fraction of the computational cost. Therefore, δ-robust AXps offer an attractive trade-off between runtime and the size of the bounds. This might be appealing in real-time decision support systems or in applications with limited computational resources, where they aim to find explanations for a large, but not necessarily maximal, feature intervals. Note that the coverage of δ-robust AXps increases monotonically with δ across all datasets without any computational overhead. One can thus disregard the computational cost in choosing δ and instead choose it according to the needs of the application in question, or to maximize the are covered subject to there still existing AXps.

4.2 Evaluation of Group Explanations

To the best of our knowledge, there is no standard dataset or approach to evaluate group abductive explanations, so we evaluate the quality of δ-gAXps with the so-called weighted FFA (Yu et al., 2023), which has been introduced to offer a formally guaranteed alternative to heuristic feature attribution methods, such as SHAP. The weighted FFA score of a feature for a sample is, intuitively, the proportion of AXps in which this feature occurs in, weighted so that shorter explanations have additional importance. To compute the FFA, we generate all AXps with the MARCO algorithm (Liffiton et al., 2015; Ignatiev & Marques-Silva, 2021) and compute the weighted FFA (Yu et al., 2023). We compare the FFA scores generated using δ-gAXps to the FFA scores from each individual sample in the group, and evaluate how close they are.

To put the result in context, we provide the same evaluation for three further approaches: "Aggregated group FFA", SHAP, and LIME. The Aggregated group FFA is the brute-force approach of computing all AXps for each sample separately and taking the mean of the FFA score for each feature. As metrics, we first use the *feature importance error*, which is the average normalized mean absolute error between the group-FFA and the FFA of each sample in the group. We provide a scale-invariant measure of accuracy by normalizing the mean absolute error by the mean of the target feature importance values. Secondly, we evaluate the *feature importance correlation* as the average Spearman correlation between the ranking induced by the group-FFA and the ranking induced by

	Diabetes		Glass		Leaf		Parkinsons		Wine		Synthetic	
	Err.	**Cor.**	**Err.**	**Cor.**	**Err.**	**Cor.**	**Err.**	**Cor.**	**Err.**	**Cor.**	**Err.**	**Cor.**
Aggr.	**.232**	**.727**	**.202**	.563	**.095**	.817	**.109**	.907	**.114**	**.852**	**.176**	**.746**
LIME	–	.493	–	.164	–	.200	–	.482	–	.586	–	.529
SHAP	–	.588	–	.355	–	.374	–	.592	–	.525	–	.634
$\delta = .00$	.277	.684	.279	.712	**.098**	**.855**	.117	.905	**.130**	**.824**	**.239**	**.670**
$\delta = .02$	**.271**	.699	.288	.712	.105	.840	**.116**	.907	.133	.818	.240	.661
$\delta = .04$	.272	.716	.277	.706	.110	.828	.117	.906	.135	.813	.246	.657
$\delta = .06$	.273	**.723**	.280	.725	.106	.832	.118	.906	.135	.811	.250	.659
$\delta = .08$	.274	.712	.271	**.746**	.105	.837	.118	.906	.135	.810	.254	.664
$\delta = .10$	.276	.708	.258	.738	.106	.827	.119	.907	.138	.817	.256	.664
$\delta = .12$	.292	.695	.258	.738	.107	.826	.119	.907	.138	.819	.257	.655
$\delta = .14$	.284	.720	**.250**	.739	.111	.828	.121	.908	.134	.820	.259	.655
$\delta = .16$	.286	.702	.253	.737	.112	.829	.117	**.910**	.134	.817	.260	.652
$\delta = .18$	.278	.701	.265	.729	.111	.818	.121	.902	.137	.821	.261	.640

Table 1: Average feature importance error (Err.) and ranking correlation (Cor.) across 10 different groups for aggregated explanations (Aggr.), SHAP, LIME, and δ-gAXps for $\delta \in \{.00, .02, ..., .18\}$. We denote in bold the best scores overall and among δ-gAXps.

the FFA of each sample in the group. We compute the ranking by ordering features according to their FFA score. We also compare our approach to the heuristic explanation methods SHAP and LIME by ordering the features in the order of importance given by SHAP and LIME and computing the correlation between their order and the ranking induced by the FFA of each sample. For SHAP and LIME, we do not compare the feature importance scores given by their FFA since the scores themselves are not commensurable.

In Table 1 we see that δ-gAXps are more accurate than SHAP and LIME. The FFA from δ-gAXps have at least 10 percentage points higher correlation to the ground-truth than SHAP on every dataset except the synthetic, and on some datasets more than 30 percentage points higher, while they have 10 to 40 percentage points higher correlation than LIME. The brute-force aggregation approach generally produces a lower error and higher correlation than δ-gAXp, but in most cases not by much. In some cases δ-gAXps produce *higher* correlation than the aggregation approach, which is notable since the aggregation considers (and thus has to compute) all explanations for all samples separately, while computing δ-gAXps is much more tractable.

Figure 5 illustrates the computational advantage of δ-gAXps. We report three runtime measurements. 'Individual' refers to the mean runtime for computing all AXps for each individual sample, 'Group' refers to the mean runtime for computing all δ-gAXps across all values of δ (as listed in Table 1), and 'Aggregated' refers to the mean runtime for the naive aggregation method that computes all AXps for all samples in a group. In the plot on the left, we see that δ-gAXps are an order of magnitude faster to compute than aggregated explanations, and that, as the number of features increases, the runtime of the naive aggregation method increases faster than the runtime of δ-gAXps. We also see that the runtime of δ-gAXps is similar to the runtime of computing all AXps for a single sample (compare Group to Individual). In the plot on the right, we see that the runtime of δ-gAXps remains constant as the group size increases, while the runtime of aggregated explanations increases linearly. Regarding the runtime of computing AXps for individual samples (right side of Figure 5), the mean runtime is 183 seconds and the standard deviation is 1384 seconds (note that these are independent of the group size). This points to another benefit of δ-gAXps: they seem to avoid some outliers with a huge number of AXps that make computing all AXps for some samples much slower, as can be seen by the fact that computing all δ-gAXps takes well below 100 seconds for each group size.

Summary of the evaluation δ-robust AXps achieve up to 20–50% of the coverage of inflated abductive explanations and are 4 to 30 times faster to compute. The FFA scores induced by δ-gAXps are comparable to (and sometimes better than) the FFA scores induced by the aggregated method, while being an order of magnitude faster to compute. Furthermore, the runtime of δ-gAXps is independent of the size of the groups of samples, and it scales similarly to the runtime of individual explanations with respect to dataset dimensionality.

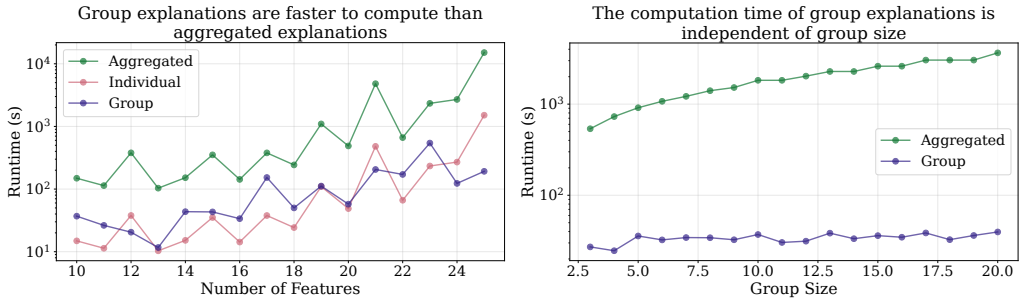


Figure 5: Mean time (s) to generate explanations for (left) groups of 10 samples on datasets with different numbers of features (10–25) and (right) groups of different sizes (3–20 samples) on datasets with 10 features. Each point represents the average of 25 instances (5 groups · 5 datasets).

4.3 Limitations

A limitation of δ-robust AXps is a trade-off between δ and the number of explanations: there might be only a few or no explanations at all, especially with larger values for δ. For example, in Figure 4 (left) less than 50% of samples have an explanation for some datasets with $\delta = 0.03$. Additionally, the possible bounds for different features might be lopsided, and in that case finding close to maximal values for δ might be difficult and one would need to revert to computing inflated AXps.

A limitation of δ-gAXps is that, in our experiments, δ-gAXps are not formally valid for each individual sample in the group. This is because the features within a group vary by more than the values of δ under which some formal explanations can still be found. This might limit the usefulness of δ-gAXps in some applications, although our evaluation based on FFA values shows that the given explanations correctly identify features that are important for the obtained classification of the samples in a group. To overcome this, one could divide groups into smaller subgroups of more similar samples, or even combining δ-robust AXps with a distance restriction on the irrelevant features (Izza et al., 2024a), since the lower the distance restriction, the easier it is to find an explanation.

Finally, in our evaluation we compute a single inflated AXp among multiple possible ones, so an inflated AXp with more volume might exist. Very recently it was shown that computing a maximum-size inflated AXp is much more computationally expensive than computing an arbitrary inflated AXp (Izza et al., 2025), so δ-robust AXps would cover less area compared to a maximum inflated AXp, but they would be even faster.

5 Conclusion

We introduced δ-robust AXps, extending robustness analysis from counterfactual to abductive explanations in a way that complements existing robustness notions for abductive explanations such as inflated and distance-restricted abductive explanations. The notion of δ-robust AXp defines formal abductive explanations that are robust to input variations, making them more practical for real-world applications with measurement noise, such as healthcare, finance, and autonomous systems. Furthermore, we introduced δ-gAXps, semi-local explanations that explain groups of samples instead of individual samples. Our empirical evaluation demonstrated that δ-robust AXps achieve 20–50% of the area coverage of inflated abductive explanations while requiring 4 to 30 times less runtime. We also demonstrated that δ-gAXps capture commonly important features within groups of samples, achieving an accuracy comparable to a brute-force method that aggregates all explanations for all samples in the group separately, but at a fraction of the computational cost.

Reproducibility Statement

All code related to the presented experiments is provided in the supplementary material with a README file describing how to rerun the experiments.

REFERENCES

Stefan Aeberhard and Michele Forina. Wine dataset, 1992. URL https://doi.org/10.24432/C5PC7J.

Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. Computing abductive explanations for boosted trees. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4699–4711. PMLR, 2023.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Min. Knowl. Discov.*, 37(5):1719–1778, 2023.

Emilio Carrizosa, Jasone Ramírez-Ayerbe, and Dolores Romero Morales. Mathematical optimization modelling for group counterfactual explanations. *Eur. J. Oper. Res.*, 319(2):399–412, 2024a.

Emilio Carrizosa, Jasone Ramírez-Ayerbe, and Dolores Romero Morales. Generating collective counterfactual explanations in score-based classification via mathematical optimization. *Expert Syst. Appl.*, 238(Part E):121954, 2024b. doi: 10.1016/J.ESWA.2023.121954. URL https://doi.org/10.1016/j.eswa.2023.121954.

John W. Chinneck and Erik W. Dravnieks. Locating minimal infeasible constraint sets in linear programs. *INFORMS J. Comput.*, 3(2):157–168, 1991.

Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. A historical perspective of explainable artificial intelligence. *WIREs Data Mining Knowl. Discov.*, 11(1), 2021.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.*, 38(5):2770–2824, 2024.

Xuanxiang Huang and João Marques-Silva. The inadequacy of shapley values for explainability. *CoRR*, abs/2302.08160, 2023a.

Xuanxiang Huang and João Marques-Silva. From robustness to explainability and back again. *CoRR*, abs/2306.03048, 2023b. doi: 10.48550/ARXIV.2306.03048. URL https://doi.org/10.48550/arXiv.2306.03048.

Alexey Ignatiev and João Marques-Silva. Sat-based rigorous explanations for decision lists. In Chu-Min Li and Felip Manyà (eds.), *Proceedings of Theory and Applications of Satisfiability Testing*, volume 12831 of *Lecture Notes in Computer Science*, pp. 251–269. Springer, 2021.

Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence*, pp. 1511–1519. AAAI Press, 2019.

Yacine Izza and João Marques-Silva. On explaining random forests with SAT. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2584–2591. ijcai.org, 2021.

Yacine Izza, Xuanxiang Huang, António Morgado, Jordi Planes, Alexey Ignatiev, and João Marques-Silva. Distance-restricted explanations: Theoretical underpinnings & efficient implementation. In Pierre Marquis, Magdalena Ortiz, and Maurice Pagnucco (eds.), *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024*, 2024a.

Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and Joao Marques-Silva. Delivering inflated explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12744–12753, mar 2024b. ISSN 2159-5399. doi: 10.1609/aaai.v38i11.29170. URL http://dx.doi.org/10.1609/aaai.v38i11.29170.

Yacine Izza, Alexey Ignatiev, Sasha Rubin, João Marques-Silva, and Peter J. Stuckey. Most general explanations of tree ensembles (extended version). *CoRR*, abs/2505.10991, 2025. doi: 10.48550/ARXIV.2505.10991. URL https://doi.org/10.48550/arXiv.2505.10991.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 14901–14909. AAAI Press, 2023.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 8086–8094. ijcai.org, 2024.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 895–905. PMLR, 2020.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5):95:1–95:29, 2023.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5491–5500. PMLR, 2020.

Mark Liffiton, Alessandro Previti, Ammar Malik, and Joao Marques-Silva. Fast, flexible mus enumeration. *Constraints*, 21, 03 2015. doi: 10.1007/s10601-015-9183-0.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pp. 4765–4774, 2017.

João Marques-Silva. Logic-based explainability: Past, present and future. In Tiziana Margaria and Bernhard Steffen (eds.), *Proceedings of the 12th International Symposium on Leveraging Applications of Formal Methods Verification and Validation*, volume 15222 of *Lecture Notes in Computer Science*, pp. 181–204. Springer, 2024.

João Marques-Silva and Xuanxiang Huang. Explainability is *Not* a game. *Commun. ACM*, 67(7): 66–75, 2024.

João Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp. 12342–12350. AAAI Press, 2022.

Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. Explainable artificial intelligence: a comprehensive review. *Artif. Intell. Rev.*, 55(5):3503–3568, 2022.

OpenML. Glass dataset. https://www.openml.org/d/41, a. OpenML ID: 41.

OpenML. Leaf dataset. https://www.openml.org/d/1482, b. OpenML ID: 1482.

OpenML. Parkinsons dataset. https://www.openml.org/d/1488, c. OpenML ID: 1488.

Axel Parmentier and Thibaut Vidal. Optimal counterfactual explanations in tree ensembles. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8422–8431. PMLR, 2021.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016.

Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In Jérôme Lang (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 5103–5111. ijcai.org, 2018.

Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018.

Greta Warren, Mark T. Keane, Christophe Guéret, and Eoin Delaney. Explaining groups of instances counterfactually for XAI: A use case, algorithm and user study for group-counterfactuals. *CoRR*, abs/2303.09297, 2023. doi: 10.48550/ARXIV.2303.09297. URL https://doi.org/10.48550/arXiv.2303.09297.

Patryk Wielopolski, Oleksii Furman, Jerzy Stefanowski, and Maciej Zieba. Unifying perspectives: Plausible counterfactual explanations on global, group-wise, and local levels. *CoRR*, abs/2405.17642, 2024. doi: 10.48550/ARXIV.2405.17642. URL https://doi.org/10.48550/arXiv.2405.17642.

Haoze Wu, Omri Isac, Aleksandar Zeljic, Teruhiro Tagomori, Matthew L. Daggitt, Wen Kokke, Idan Refaeli, Guy Amir, Kyle Julian, Shahaf Bassan, Pei Huang, Ori Lahav, Min Wu, Min Zhang, Ekaterina Komendantskaya, Guy Katz, and Clark W. Barrett. Marabou 2.0: A versatile formal analyzer of neural networks. In Arie Gurfinkel and Vijay Ganesh (eds.), *Computer Aided Verification - 36th International Conference, CAV*, volume 14682 of *Lecture Notes in Computer Science*, pp. 249–264. Springer, 2024.

Min Wu, Haoze Wu, and Clark Barrett. Verix: Towards verified explainability of deep neural networks, 2023a. URL https://arxiv.org/abs/2212.01051.

Min Wu, Haoze Wu, and Clark W. Barrett. Verix: Towards verified explainability of deep neural networks. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, 2023b.

Jinqiang Yu, Alexey Ignatiev, and Peter J. Stuckey. On formal feature attribution and its approximation, 2023. URL https://arxiv.org/abs/2307.03380.

A Appendix

(Wu et al., 2023a) proposed a sensitivity-based ordering to compute small AXps. The idea of their approach is to sort features by how much they influence the output of the classifier. The feature that influences the output the least is the first feature in the ordering. This approach leads to small explanations since it is more probable that features to which the classifier is not sensitive are irrelevant and can thus be removed from the over-approximation of an AXp. We experiment with a smoothed version of the sensitivity of a feature by averaging it with the sensitivity of its neighbors, in order to produce explanations that include pixels that are close to each other. We call our approach smoothed sensitivity ordering. In Figure 6 we can see the explanations generated using different orderings. The VeriX ordering produces explanations that are small but are unintuitive, while the smoothed sensitivity ordering produces explanations that are concentrated towards the center and capture the shape of the digit better.

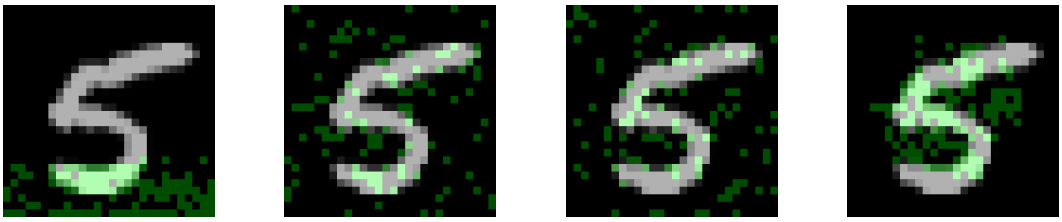


Figure 6: The first explanation generated using different orderings. From left to right, the orderings are: linear, random, VeriX, and smoothed sensitivity. The green pixels represent the relevant features in the explanations.