POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

AXOM: Combination of Weak Learners' eXplanations to Improve Robustness of Ensemble's eXplanations

Laurea Magistrale in Computer Engineering - Ingegneria Informatica

Author: Riccardo Pala

Advisor: Prof. Daniele Loiacono

Co-advisor: Esteban García-Cuesta

Academic year: 2022-2023

1. Introduction

1.1. Rationale

Machine learning is increasingly becoming an important aspect of various fields in which decisions can heavily affect humans' lives. This means that, as machine learning models are deployed in real-world applications, it becomes crucial to ensure their trustworthiness and reliability. One way to achieve these important properties is through providing model explanations by means of XAI algorithms. However, recent studies have shown that, alongside the broad variety of these methods, there are still a number of concerns regarding the *robustness* of the explanation values that they are able to produce, particularly in situations in which models are feed with inputs that lay outside the data distribution they were trained on. Explanations provided by models may be sensitive to small changes in the input data, leading to unreliable explanations. All these concerns justify the need to carry out investigations in order to develop explanation methods capable of improve the robustness.

1.2. State of the Art

In many areas of competence the problem of *explaining* models' outputs may become a matter of great importance [7]. When a model does not meet the requirements to be considered intrinsically explainable, it is necessary to apply post-hoc eXplainable AI (XAI) techniques which can be divided into *model-agnostic* or *model-specific*. One of the most widely used is *SHAP* (SHapley Additive exPlanations) [10], an XAI method to explain individual predictions, which calculates the contributions of each features as its average marginal contribution across all possible partitions of the feature space. Although this XAI technique enjoys many desirable properties, it also suffers from several limitations, mainly concerning its robustness. Intuitively, robustness means that similar inputs must produce similar explanations [2]. Ensembles of models have some desirable properties to provide more accurate and robust predictions, increasing the generalization capabilities of the single models by training several of them and combining their decisions to obtain a single prediction [5]. Particular attention is placed on *Random Forest* (RF), a very successful model that combines independent *Decision Tree* (DT) models to build up a

more powerful learner. Each tree casts a unit vote for the most popular class to assign to the input sample. This results in an increment in the classification accuracy as well as in the ability to generalize [3, 8]. Concerning the intrinsic explainability of this technique, however, this decreases as the number of weak learners in the ensemble increases. One way to overcome this problem is to apply XAI model-agnostic algorithms, such as SHAP. At [2] the authors observed that such an application yields values that only partially meet the expectations. Although the method provides the contribution of every feature, in most cases the corresponding values have low robustness to small perturbations of the input. This can be interpreted as a symptom of a lack of trustworthiness of these explanations.

1.3. Objectives

This work is devoted to the conduction of a study aimed at the developing of effective methods for the application of model-agnostic XAI techniques to model ensembles. The goal is to find a way to exploit the excellent prediction capabilities and improved robustness of this category of models in order to enable the production of more robust explanation values. What makes this research interesting is the promise that such application, by paralleling the already established improvements in predictive capabilities, is able to make explanations more robust and therefore trustable. What we are aiming to is the production by the developed procedure of explanation values that result more robust to small deviations in the input. This means that, given a certain data point x and a slightly perturbed version of it x', we expect the explanations y and y', respectively produced from the two inputs just mentioned, to differ marginally.

2. Development

2.1. Robustness Metric Definition

As a first step, we need to rigorously define a method to quantify the robustness property of an explanation. The choice, basing on the work presented in [2], fell on the notion of Lipschitz continuity, defined as:

Definition 2.1. *$f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is **locally Lipschitz** if for every x_0 there exist $\delta > 0$ and $L \in \mathbb{R}$ such that $||x - x_0|| < \delta$ implies $||f(x) - f(x_0)|| \leq L||x - x_0||$.*

Making use of this notion, we introduce the robustness criteria to calculate the average value of the incremental ratio around the data point of interest. Let $\mathcal{X}$ be the input space, let $\mathcal{A}$ be the set of features of $\mathcal{X}$ and let $f(\cdot)$ be the prediction function of the model. Define, for every x_i sample of the test set, a discriminative discretization of its surrounding:

$$\mathcal{N}_{f,\epsilon}(x_i) = \{x_j \in \mathcal{X} \mid |x_{i,a} - x_{j,a}| \leq \epsilon \ \forall a \in \mathcal{A},\ f(x_i) = f(x_j)\}$$

while we now want to calculate the robustness of the SHAP explanations on data point x_i with:

$$\bar{L}(x_i) = \frac{1}{|\mathcal{N}_{f,\epsilon}(x_i)|} \sum_{x_j \in \mathcal{N}_{f,\epsilon}(x_i)} \frac{||g(x_i) - g(x_j)||_2}{||x_i - x_j||_2}$$

where $g(\cdot)$ is the SHAP explanation function. We decided to include in the robustness calculation only the perturbed samples whose label predicted by the model is the same as the original sample, because we expect that a perturbed data point whose label differs from that of the original data point will produce an explanation that differs substantially from that of the original data point, since different outputs are understood with different explanations.

2.2. Combination of Weak eXplanations

DT is a model with very good intrinsic explainability but by design it creates hard decision boundaries meaning that small changes in the input can lead to abrupt changes in the explanations. Indeed, explanation values are constant within the *zones of explanation constancy*, i.e. the portions of space in which predictions and explanations comes from the activation of the same branch, which can be identified as the cause of the abrupt changes in the SHAP explanations of DT. RF, on the other hand, relies on the combination of several weak learners to create smoother SHAP explanation boundaries coming from the averaging of the weak ones. What we expect is that the softer boundaries provide more robust explanations. Fig. 1 shows a 2-dimensional comparison between DT and RF explanations, being x_i the central point and with

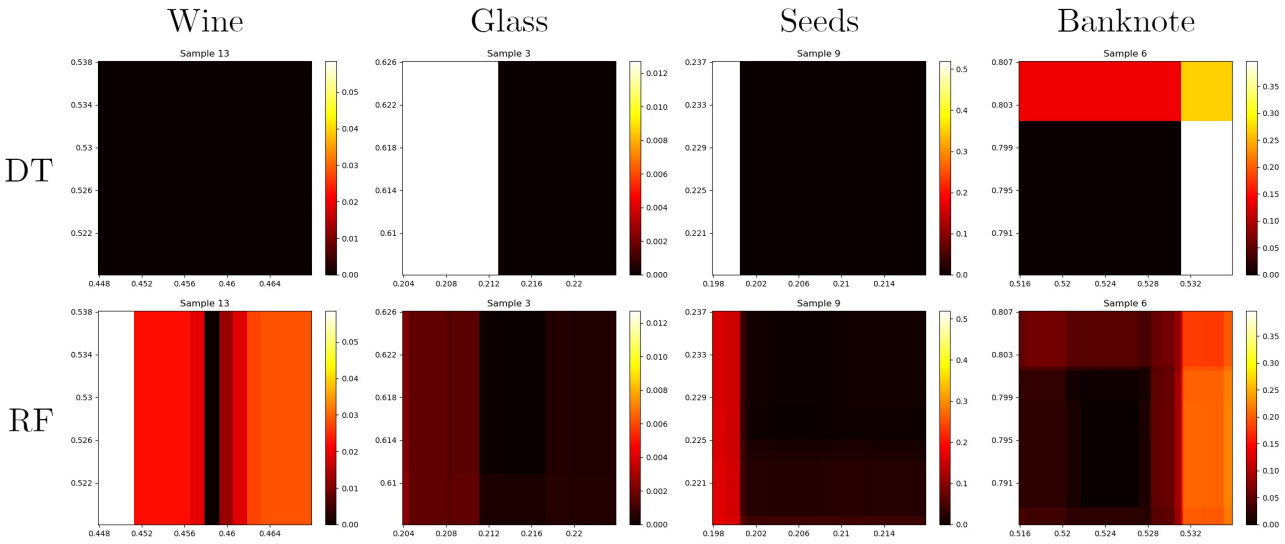

Figure 1: Comparison between DT and RF regarding differences in the explanations values around some x_i samples.

each value $H_d(x_j)$ of the map computed through the formula:

$$H_d(x_j) = ||g(x_i) - g(x_j)||_2$$

As we expected, RF heatmaps depict more gradual color changes, which correspond to differences in values that follow a smoother progression. However, while changes in RF explanations are smoother, we will see that they hardly provide improvements on the robustness (see Table 2), probably because complex ensembles, though often better performing, tend to be more sensitive to the noise generated by the explanations of the weak learners that provided a wrong label. For this reason, it may be logical to "reward" those explanations from the models that contributed positively to the final decision.

2.3. Averaging on the eXplanations Of the Majority (AXOM)

For our solution, we propose to combine the explanations only of weak learners which prediction matches with the one obtained by the ensemble (see Fig. 2 for a more clear graphical explanation). In algorithm 1 the AXOM evaluation algorithm for each data point x is presented. The method receives as parameters the ensemble e (a RF trained model), the data point x and the SHAP explainer σ. $\phi_w \in \mathbb{R}^{1\times p}$ contains the explanations of the weak learner, being p the number of features, and an explanation is added to $\Phi \in \mathbb{R}^{n\times p}$, being n the number of selected weak learners, if the label provided by the weak learner l_w is equal to that predicted by the ensemble l_e. The final explanation $axom\_shap \in \mathbb{R}^{1\times p}$ is the mean of all selected weak explanations Φ for sample x.

Algorithm 1 AXOM procedure to calculate single-sample explanations for an ensemble

procedure AXOM_SHAP_EXP(e, x, σ)
  $l_e \leftarrow e$.PREDICT(x)
  $W \leftarrow e$.estimators
  $\Phi \leftarrow$ **new** LIST( )
  for w **in** W **do**
    $l_w \leftarrow w$.PREDICT(x)
    if $l_w = l_e$ **then**
      $\phi \leftarrow$SHAP_EXPLANATION(w, x, σ)
      Φ.APPEND(ϕ)
    end if
  end for
  $axom\_shap \leftarrow \frac{1}{|\Phi|}\sum_{\phi\in\Phi}\phi$
  return $axom\_shap$
end procedure

This method ensures that the obtained explanation is free from the noise resulting from the explanations of the weak learners that provided a different label from the ensemble. Arguably, this improves the quality of the explanation only in the case where the ensemble has provided correct output, so we expect to obtain data that support the decision that was actually made, extractable from the majority, which makes the method useful for understanding what led to the decision.

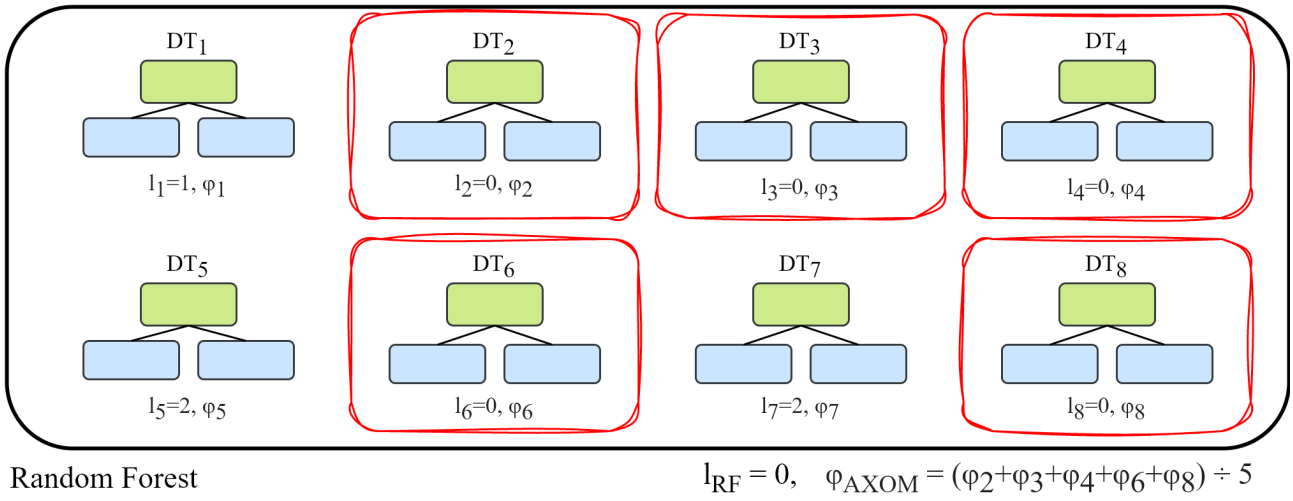

Figure 2: A toy example is depicted in the figure to illustrate the functioning of AXOM. The method considers in the average only the explanations of the weak learners who were part of the majority in the voting.

2.4. Datasets

We tested the methods on four commonly used datasets from **UCI Machine Learning Repository**, that is Wine [1], Glass Identification [6], Seeds [4], Banknote Authentication [9], being the test sets with a size of 10% of total data. In table 1 we specify some information and the accuracy of the tested models. All of these address a classification task with multivariate data points. The number of features of the tested datasets was limited to $\lfloor \log_2(10000) \rfloor = 13$, given the choice of using 10000 points to evaluate the robustness within the neighbourhood, in order to guarantee an adequate search in the entire feature space, that is, with at least two perturbations along each feature axis.

Datasets	N. features	N. samples	Accuracy	
			DT	RF
Wine	13	178	88.9%	100%
Glass	10	214	81.8%	95.5%
Seeds	7	210	85.7%	95.2%
Banknote	5	1372	98.6%	99.3%

Table 1: Descriptions of the datasets.

2.5. Experimental Design

With regard to the conducted tests on robustness, the choice of the radius of the neighbourhood of the data points of the test set to be analysed was $\epsilon = 0.01$. This value defines the perturbation area to be analyzed and it is constant for all the experiments, in which data samples were standardized to 0-1. The same experiments were done for DT and RF. Two functions were then defined for calculating the value of $\bar{L}$, one that performs the calculation through the explanations obtained directly from the models and one that performs it on RF through AXOM algorithm. For each data point x_i of the different test sets, 10000 perturbed x_j samples are randomly chosen from their relative ϵ-neighborhood $\mathcal{N}_\epsilon(x_i)$, on which the variation of the explanation value is calculated through the formula in section 2.1, *if and only if its corresponding predicted label is equal to the one predicted for* x_i. This latter choice derives from the fact that we only expect robust explanation values as long as these only account for a single output value. Indeed, we expect that a perturbed data point whose label differs from that of the original data point will produce an explanation that differs substantially from that of the original data point, since different outputs are understood with different explanations.

3. Results

Table 2 shows the $\bar{L}$ results in the form of mean and standard deviation for each model and dataset, while Fig. 3 present them in a more detailed way by means of box plots (**Note**: $\bar{L}$ indicates the variation of explanation values in the neighborhood, thus a lower mean value is associated with a higher robustness). Mean and standard deviation values of AXOM are better in each of the four analyzed datasets compared with RF. However, when comparing the robust-

ness values of AXOM with those of DT, the former provides significant improvements only in some datasets. In the later parts of this chapter we will analyze in detail the reasons for this anomalous behavior.
To verify the reliability of these results, two-sample t-test was used. As we expected, Table 3 shows that AXOM significantly improves robustness over RF for all datasets (see row RF vs. AXOM), with p-values all below the 0.05 threshold. Regarding the DT vs. AXOM comparison, neglecting for the moment the case of the WINE dataset to be discussed, it can be observed that the equality in mean robustness in SEEDS is reflected in a statistical improvement in favor of AXOM, which possesses values deviating from the mean with less magnitude (also observable in Fig. 3).

Comparison	p values			
	WINE	GLASS	SEEDS	BANK
DT vs. RF	<0.001	0.774	0.008	0.3023
DT vs. AXOM	<0.001	0.113	0.008	0.0656
RF vs. AXOM	0.042	0.007	0.030	0.0444

Table 3: Two samples mean T-test values comparison.

To get some insights, Fig. 4 shows the 2-D robustness heatmaps for the four test sets, constructed by averaging the heatmaps of all test samples. Being $\mathcal{T}$ the set of all test points of a given dataset, all $x_i \in \mathcal{T}$ test samples were centered in (0, 0) so that all samples could be fit in the same box with axes bounded inside $(-\epsilon, \epsilon)$, while the *heat* values $H(x_j)$ are computed through:

$$H(x_j) = \frac{1}{|\mathcal{T}|} \sum_{x_i \in \mathcal{T}} \frac{||g(x_i) - g(x_j)||_2}{||x_i - x_j||_2}$$

It is possible to see from the colors of the plots that RF and AXOM (except, again, for WINE dataset) always exhibit more desirable behavior than DT, with significantly smaller explanation values that vary considerably more smoothly. Comparing RF vs. AXOM we can notice a very similar behavior, with the only difference represented by the lower values of the latter. Obviously, in order to make possible the 2-D representation, this graphical analysis was conducted by constructing the surroundings with perturbations only along two arbitrary axes of the feature space.
About the anomalous robustness values of DT, first it is worth to recall that the accuracy of the RF model is better than DT (see Table 1), and so also the expected quality of the explanations. Having said that, AXOM improves robustness for all datasets except for WINE dataset, when compared with DT. This is due a fortuitous behavior of DT for the experimental design parameters. Indeed, it can be observed from Table 2 that the obtained robustness is 0. That would mean that the robustness is "perfect", i.e. all the SHAP explanations for the perturbed data have exactly the same value as the original data point.

Model	WINE		GLASS		SEEDS		BANKNOTE	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Decision Tree	0.00	0.00	1.89	1.78	0.65	2.02	1.75	3.32
Random Forest	0.55	0.51	1.75	1.87	0.77	0.72	1.58	1.57
AXOM	0.47	0.44	1.27	0.72	0.65	0.67	1.28	1.34

Table 2: Mean and standard deviation of the $\bar{L}(x_i)$ values calculated for each sample x_i of the various test sets. Lower values of $\bar{L}$ denote better robustness to perturbations.

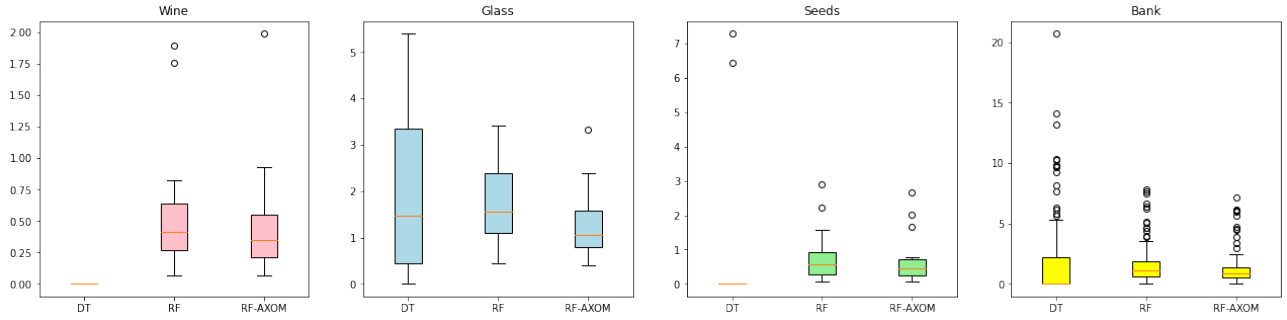

Figure 3: Box plots constructed from the $\bar{L}(x_i)$ values of the x_i samples of the test sets of each of the four analysed datasets.

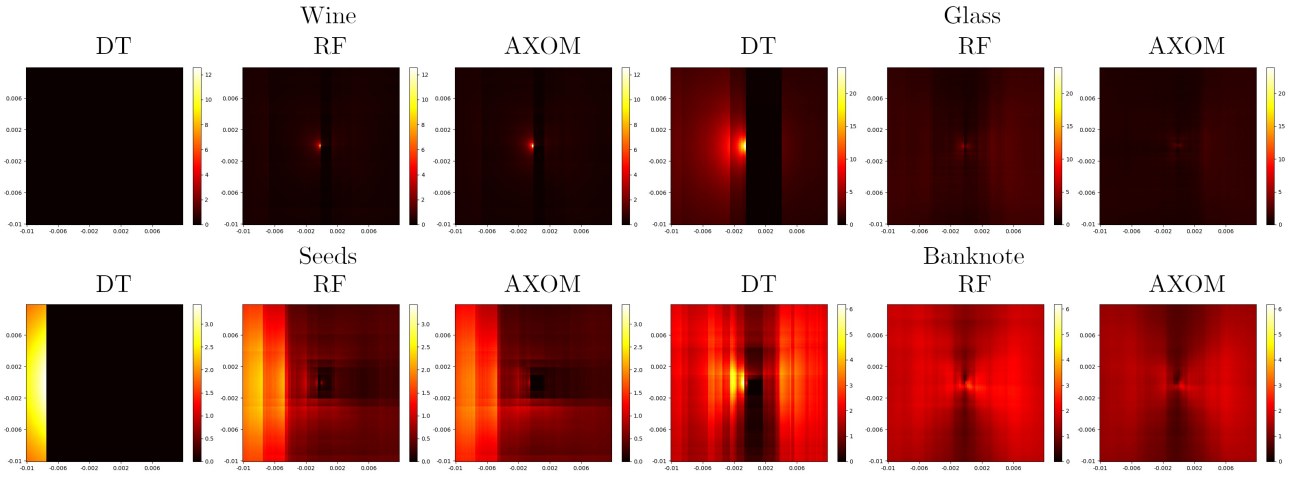

Figure 4: Comparison of the $\bar{L}$ heatmaps (two features) of the explanations for DT, RF and AXOM for the entire dataset.

However, the production of explanations that are constant over a large portion of the feature space is far from being a desirable behaviour, especially as the complexity of the problem increases. Differences in explanations of DT are zero-valued when all the perturbed points activate the same decision branch of the sample under analysis. DT provides explanations that enjoy perfect robustness only in samples that are sufficiently distant from the decision surfaces where a branch change occurs, whereas for data points that are more *unlucky* in this respect, the value of $\bar{L}$ is significantly larger. An example of this behaviour is shown in Fig. 5 through two representative samples of Wine test set. Specifically, by setting $\epsilon = 0.2$ DT provides explanations that suffer abrupt changes in value as soon as a change in the decision branch is reached, with values significantly higher than those of AXOM. Note that in DT a change of decision branch does not necessarily imply a change of predicted label.

4. Conclusions

By analysing the elements on which trustworthiness of explanations is based, robustness was identified as a pivotal property. In this work, we presented as a solution the establishment of an unambiguous criterion for measuring such quality and a procedure for calculating SHAP explanations of ensembles as the result of averaging explanations of weak learners who contributed positively to the final prediction. This approach has proven to significantly improve the robustness, confirming that weak learners taken individually can play a key role in explaining the decisions of the ensemble to which they belong. In particular, a discriminative combination enables the reduction of variance in the explanations eliminating noise deriving from the weak learners who provided an incorrect prediction. We envisage that this approach is not limited to RF and SHAP and that it is natural to extend it to other types of ensembles, such as Bagging or Gradient Boosting, as well as to other post-hoc XAI techniques.

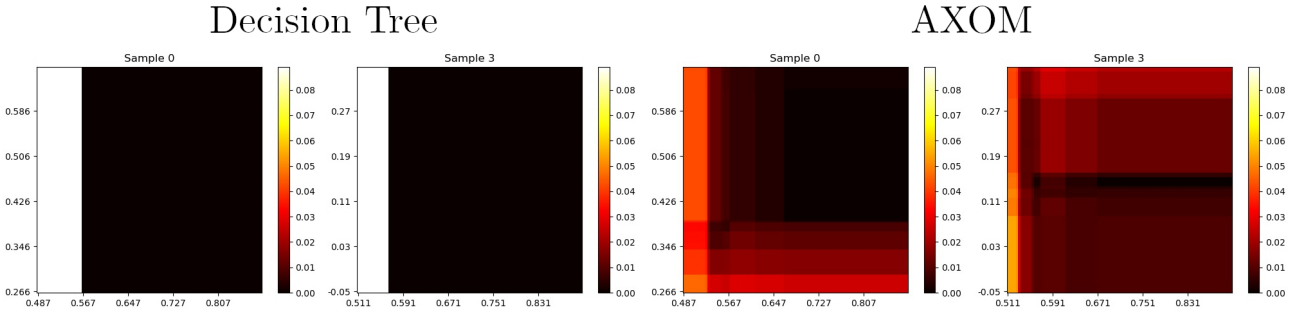

Figure 5: Comparison of DT and AXOM explanations difference heatmaps for two representative samples of Wine dataset with $\epsilon = 0.2$.

References

[1] S. Aeberhard. Wine. UCI Machine Learning Repository, 1991.

[2] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Magorzata Charytanowicz, Jerzy Niewczas, Piotr Kulczycki, Piotr Kowalski, and Szymon Lukasik. seeds. UCI Machine Learning Repository, 2012.

[5] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012.

[6] B. German. Glass Identification. UCI Machine Learning Repository, 1987.

[7] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 10 2017.

[8] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995.

[9] V. Lohweg. banknote authentication. UCI Machine Learning Repository, 2013.

[10] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.