Chapter 5

Comparisons

5.1 Introduction

One will follow and define the theoretical framework laid by D. Canha et al (2025) [14]. It is one of the most recent and well constructed framework up to date in the field. They use 11 mathematically defined properties, and their sub-properties, to have a general picture of the strength and weaknesses of the XAI method. The strength of a functionally-grounded evaluation is that it does not require human evaluation and can be automated. It offers a common ground onto which to compare different XAI methods. It is inspired by the Ethics Guidelines for Trustworthy AI published by the European Union's High Level Expert Group (AI HLEG) [20]. The guidelines identify 3 axes participating to the transparency of an AI system :

- Explainability : how the model works, how it makes decisions.
- Traceability : how data and labels have been gathered, what classifier has been used, etc...
- Communication : how explanations can be adapted to the explainee depending on his/her social situation.

In the following analysis, the results are to be understand as what they are : a comparison between XAI methods belonging to different model's families, under a benchmark framework primarily thought for post-hoc attribution based methods, in the context of this classification task on (mostly) categorical data and not-so-accurate predictions models. Some methods will be advantaged or disadvantaged depending on their nature and the context of the evaluation.
Therefore, all results should be interpreted with caution, not as ground truth, and taking into the aforementioned context.
The results in sections "Implementation and Results" are presented in the following way "**method name** (*number*)". The *number* represents the score achieved for the sub-metric. For instance, "**DiCE** (3)" corresponds to a score of 3 for a given metric.

5.2 F1 - Representativeness

5.2.1 F1: Normative definition

Representativeness is about how deeply the XAI method describes the internal workings of the black-box and what type of model and input data it supports. It is broken down into 4 sub-properties denoted by F1.1 to F1.4.

Scope (F1.1)

This property evaluates whether the method give local or global explanations (the entire model behavior vs a single prediction). D. Canha et al (2025) claims that Local Explanations are preferred over Global ones, which explains the following scoring.

$$m_{f1.1} = \begin{cases} 3 & \text{the XAI method supports both Global and Local expl.,} \\ 2 & \text{the XAI method is local} \\ 1 & \text{the XAI method is Global} \end{cases} \tag{5.1}$$

Portability (F1.2)

This property addresses the range of the ML models to which the XAI method can be applied : either *model-specific* methods or *model-agnostic* methods. The former can only be used on certain kind of model, while the latter can be applied to any kind of models. The higher scores are assigned to model-agnostic methods because of their generability.

$$m_{f1.2} = \begin{cases} 2 & \text{the XAI method is model-agnostic} \\ 1 & \text{the XAI method is model-specific} \end{cases} \tag{5.2}$$

Access (F1.3)

This property's role is to evaluate the degree to which the XAI method requires access to the training data or the model itself to generate an explanation.
The metric is divided into two sub-properties : Data Access and Model Access. Together, they quantify the level of privacy-preserving capability offered by the XAI method.

$$m_{f1.3A} = \begin{cases} 3 & \text{no (training) data access required} \\ 2 & \text{data required only for initialization} \\ & \text{(e.g. creating an explainer object)} \\ 1 & \text{data required for each explanation} \\ 0 & \text{(full dataset required for (re)training)} \end{cases} \quad m_{f1.3B} = \begin{cases} 3 & \text{no model required (any function } f(X)) \\ 2 & \text{prediction function access only} \\ & \text{queries to a trained model} \\ 1 & \text{partial access (e.g. gradients)} \end{cases} \tag{5.3}$$

Practicality (F1.4)

This property measures the applicability across diverse data types and the efficiency of handling large datasets by the XAI method. It is divided into two sub-properties : $m_{f1.4A}$ Applicability and $m_{f1.4B}$ Scalability.

$$m_{f1.4A} = \begin{cases} 2 & \text{data-agnostic} \\ 1 & \text{partially data-specific} \\ & \text{(requires preprocessing)} \\ 0 & \text{fully data-specific} \end{cases} \qquad m_{f1.4B} = \begin{cases} 2 & \text{highly scalable} \\ 1 & \text{moderately scalable} \\ 0 & \text{not scalable} \end{cases} \tag{5.4}$$

5.2.2 F1 : Implementation and results

F1.1 Scope

LIME (2) Lime is a local explanation technique by definition and does not provide global explanations.

KernelSHAP (2) KernelSHAP produces local explanation ; KernelSHAP cannot give global explanations : approximated Shapley values, highly volatile and coefficients from different regressions cannot be meaningfully aggregated. Final score : 2

DiCE (2) DiCE generates instance-level counterfactual explanations and is inherently local.

OSDT (3) OSDT's learned decision tree constitutes a complete global explanation, and each prediction is explained locally by a decision path from the root to a leaf.

F1.2 Portability

LIME (2) : Model-agnostic; relies only on access to a prediction function.

KernelSHAP (2) : Model-agnostic.

DiCE (2) : Formulation is model-agnostic and data-agnostic, relying only on access to a prediction function and a defined distance metric. Final score : 2

OSDT (1) : Model-specific; OSDT's explanations are inseparable from the predictive model itself.

F1.3 Access

LIME (2 + 2) : Data only required for the initialization and it requires only an access to the predictive function, through querying a trained model.

KernelSHAP (2 + 2) : Same justification as LIME.

DiCE (2 + 2) : Same justification as LIME.

OSDT (3 + 0) : Its explanations do not require access to any external predictive model, as the explanatory structure is identical to the classifier itself. However, OSDT requires the full access to the training data to construct the explanatory model.

F1.4 Practicality

LIME (2 + 1) : Data agnostic, and moderately scalable. It can be computationnaly expensive for large datasets.

KernelSHAP (2 + 0) : Data agnostic, but impractical for large datasets.

DiCE (2 + 1) : It is most commonly applied to tabular data in practice, but the formulation is data-agnostic. DiCE can be applied to moderately large datasets.

OSDT (1 + 1) : It is primarily applicable to tabular data with finite, discretizable features. OSDT can be applied to moderately large datasets, it exhibits significant computational overhead as model complexity increases.

Summary Table below summarizes the F1 scores across all methods. We can observe that **LIME** and **DiCE** have the highest scores, indicating higher representativeness under this metrics implementation choice.

Table 5.1: F1 Practicality scores

	Scope (F1.1)	Portability (F1.2)	Access (F1.3)	Practicality (F1.4)	Total
LIME	2	2	4	3	**11**
KernelSHAP	2	2	4	2	10
DiCE	2	2	4	3	**11**
OSDT	3	1	3	2	9

5.3 F2 - Structure

This new property focus on the *how* the explanations are presented and their visual aspects.

5.3.1 F2 : Normative definition

Expressive power F2.1

This property assesses the extent to which the XAI method supports diverse and comprehensible representation formats or languages. n is the number of distinct explanatory outputs (e.g. importance scores, counterfactuals) $|F|$ is the number of unique representation formats (e.g. textual summary, visual graph of text), C is a predefined set of comprehensible formats including : decision trees, text summary, bar plots, rule sets or paths, $1(f \in C)$ is an indicator function that returns 1 if the format f belongs to the set C, 0 otherwise.

$$m_{f2.1} = n + |F| \frac{\sum_{f \in F} 1(f \in C)}{|F|} \tag{5.5}$$

In this work, Expressive Power is assessed based on the explanatory outputs explicitly produced and presented by each method, and the possibilities allowed by the explanation's format to be visually showcased.

Graphical integrity (F2.2)

This property assesses if the explanation makes a distinction between positive and negative attributions.

$$m_{F2.2} = \begin{cases} 1 & \text{pos. and neg. attributions are visually distinguishable} \\ 0 & \text{they are not visually distinguishable} \end{cases} \tag{5.6}$$

Morphological Clarity (F2.3)

This property assesses whether the explanation visually discrimates more relevant attributions from less relevant ones.

$$m_{F2.3} = \begin{cases} 1 & \text{the most relevant attributions are visually distinguisable} \\ 0 & \text{they are not visually distinguishable from less relevant ones} \end{cases} \tag{5.7}$$

Layer Separartion (F2.4)

This property assesses whether the explanation visualization omits or occludes the original input instance, which should be visible for explainee inspection.
In this work, **Layer Separation** is applied conservatively, focusing stricly on the information contained in the explanation output, not on what might be accessed outside of the explanations.

$$m_{f2.4} = \begin{cases} 1 & \text{The original input instance is visible} \\ 0 & \text{the original input instance is ommited or occluded} \end{cases} \tag{5.8}$$

5.3.2 F2 : Implementation and Results

F2.1 Expressive power

Expressive power quantifies the visual versatility of methods.

LIME (3.5) : Produces a single explanatory output : a local feature coefficients. Two representations formats are considered : table summary and a bar plot.

KernelSHAP (3.5) : Produces a single explanatory output : Shapley values. The same two representation formats as lime are considered.

DiCE (2.0) : Produces counterfactual instances as explanation and supports a single representation format : table-based summary of the counterfactual instances.

OSDT (4.0) : Produces decision paths/rules as explanations and supports two unique representation formats : rule sets and decision trees.

F2.2 Graphical integrity

LIME (1) : Positive and negative attributions are visually distinguishable in the bar plot.

KernelSHAP (1) : Same justification as LIME.

DiCE (0) : Counterfactual instances do not provide positive or negative attributions.

OSDT (0) : Same justification as DiCE.

F2.3 Morphological Clarity

LIME (1) : The bar plot visually distinguishes the most relevant attributions.

KernelSHAP (1) : Same justification as LIME.

DiCE (0) : Counterfactual instances do not visually distinguish relevant features.

OSDT (1) : Although OSDT does not assign graded attributions, its tree structure provides a *visually distinguishable* hierarchy of features.

F2.4 Layer separation

LIME (0) : The original input instance is occluded in the bar plot.

KernelSHAP (0) : Same justification as LIME.

DiCE (0) : The original input instance is not shown in the counterfactual instances.

OSDT (0) : The original input instance is occluded in the decision paths/rules.

Summary The below table summarizes the F2 scores across all sub-metrics. We can observe that **LIME** and **KernelSHAP** have the highest scores, indicating, under this metrics implementation choice, a more interpretable visual structure.

Table 5.2: F2 Structure scores

	Expr. Power (F2.1)	Graph. Int. (F2.2)	Morph. Clarity (F2.3)	Layer Sep. (F2.4)	Total
LIME	3.5	1	1	0	**5.5**
KernelSHAP	3.5	1	1	0	**5.5**
DiCE	2.0	0	0	0	2
OSDT	4.0	0	1	0	5

5.4 F3 - Selectivity

5.4.1 F3 : Normative definition

This property evaluates the length of the explanation by the methods. Based on Miller's Law, 7 ± 2 is the optimal number of information that an Human can process and remember. This property is inspired by it, reaching its maximum value when the number of explanations is 7 or it allows adjustment to $s = 7$, where s is the explanation size.

$$m_{f3} = \begin{cases} 1 & \text{if } s = 7 \text{ or tunable to } s = 7 \\ \exp(-\frac{(s-7)^2}{2\sigma^2}) & \text{if } s \neq 7 \text{and not tunable} \end{cases} \tag{5.9}$$

5.4.2 F3 : Implementation and Results

LIME (1) : The method allows one to tune the number of feature's score seen, therefore yielding a maximum score.

KernelSHAP (1) : The methods allows one to tune the number of Shapley values seen.

DiCE (0.9) : The best approximation of the explanation size s for DiCE is the average number of feature changes across multiple counterfactual, data points and target classes. The approximation is $s_{\text{DiCE}=8}$.This value is used as input to the benchmark's selectivity scoring function. However, the confidence interval of the standard deviation of this metric is $[5.5, 7.6]$. This highlights the high variability across instances and target classes in the number of feature changes required to achieve counterfactuals.

OSDT (0.6) : For OSDT, local explanations correspond to the decision paths from the root to a leaf node, leading to a prediction. Therefore, we computed the **Selectivity** metric as the mean number of conditions (feature tests) in the decision paths across all test instances. This number here is (rounded to) 5.

Summary The below table summarizes the F3 scores across all methods. We can observe that **LIME**, **KernelSHAP** and **DiCE** outperform **OSDT** in terms of quantity of information shown. It is mostly due to the computational choice of the metric that will be criticized in the discussion section. Overall, all methods perform relatively well in this metric.

Table 5.3: F3 Selectivity scores

	Total
LIME	**1.0**
KernelSHAP	**1.0**
DiCE	0.9
OSDT	0.6

5.5 F4 - Contrastivity

5.5.1 F4 : Normative definition

This property indicates whether or not the method discriminates the explanation from an alternative outcome, like counterfactual explanation would do. This idea is divided into two sub-properties : Contrastivity Level F4.1 and Target Sensitivity F4.2

Contrastivity Level (F4.1)

This sub-property assesses how effectively the generated explanation provides contrastivity relative to a standard reference point (e.g. in Shapley's value, $\phi_0 = E[f(X)]$ is used as a baseline and is the expected prediction of the model over the data distribution, which is approximate in practice by the average prediction of the model). The higher scores are given to methods that explain the model output compared to multiple reference points.

$$m_{f4.1} = \begin{cases} 2 & \text{the expl. compares the output with multiple events} \\ 1 & \text{the expl. compares the output with a single reference point} \\ 0 & \text{no Contrastivity provided} \end{cases} \tag{5.10}$$

Target Sensitivity (F4.2)

This property test the robustness of the method against adversarial attacks. A reliable explanation should be sensitive to the alteration of inputs that produce (significant) changes in the model's outputs.
Methods for generating adversarial samples differ based on data type. For our case in tabular data, we will use a nearest counterfactual approach.
E_1 represents the explanation before perturbation, E_2 represents the explanation after perturbation, $d(E_1, E_2)$ represent the distance between the two explanations (e.g. Euclidean distance for feature contributions), and d_{max} the maximum distance for normalization.

$$m_{f4.2} = \frac{d(E_1, E_2)}{d_{max}} \tag{5.11}$$

5.5.2 F4 : Implementation and Results

F4.1 Contrastivity Level : scores

LIME (1) : LIME compares the output with a single reference point, the predefined baseline, which is most commonly the locally weighted, average prediction of the black-box model in the neighborhood of the instance. The term β_0 in g_x from equation 3.1.

KernelSHAP (1) : KernelSHAP explains predictions relative to ϕ_0 in 3.2, which represents the expected model output over the background distribution.

DiCE (2) : DiCE shows by definition several multiple different CF explanations.

OSDT (0) : Although decision trees implicitly encode alternative deicison paths corresponding to different outcomes, OSDT explanations [1], following the exact wording of the sub-metric, do not explicitly articulate contrasts with a reference outcome.

F4.2 Target Sensitivity

LIME (0.3) : The euclidean distance is computed between the explanation of the original input and the explanation of the closest counterfactual to that reference input, targeting each alternative class. The closest valid counterfactual was chosen to mimic a *slight* perturbation. The sensitivity metric is summarized by the median, assuming equal relevance of alternative classes. The interquartile range is 0.1750, implying a substantial variation among classes in the explanation robustness against minimal perturbation.

KernelSHAP (0.4) : The methodology is same as for LIME. The interquartile range is 0.4000 implying a large variation.

DiCE (0.0) : The notion of target sensitivity to small perturbations of the input is unbridgeable with the design objective of the method. Consequently, this metric is not evaluated for DiCE.

OSDT (0.5) : Target-conditioned explanation generation is degenerate, since the explanation is uniquely determined by the model structure and instance. We therefore reinterpret target sensitivity as a measure of structural contrastivity, quantifying whether a predicted class is supported by a single dominant decision rationale or by multiple competing rule paths. We replace instance-level target perturbation with a global structural dispersion measure over decision paths. The sensitivity proxy is defined as $1 - p^{(c)}_{max}$, for each class c.

$$\text{sensitivity} = 1 - p^{(c)}_{max} \quad \text{where} \quad p^{(c)}_{max} = \max_{p \in \mathcal{P}_c} P(\text{path} = p \mid \hat{y} = c) \tag{5.12}$$

is the maximum conditional probability of a single decision path among all paths $\mathcal{P}_c$ predicting class c. The probability is computed as the proportion of point in that class sharing this unique path.
The final result is the median across classes of the sensitivity proxy metric to reduce sensitivity to class imbalance (i.e. outliers). Higher value of this metric, implying small p_{max}, indicates the absence of a single dominant decision path. The interquartile range is 0.4182 implying large variation among the proportion of predicted points by the dominant path by classes. It means that certain classes have one dominant decision path, i.e. mostly the same local explanation for instances, while others do not have a dominant decision path but several different ones.

[1] When we refer to OSDT's explanations, we are talking about a decision path from the root to a leaf leading to a class' prediction

Summary The table 5.4 summarize the results and showcase that, despite an absence of value for the F4.2, DiCE arises as the most contrastive method.

Table 5.4: F4 Contrastivity scores

	Contrastivity Level (F4.1)	Target Sensitivity (F4.2)	Total
LIME	1	0.3	1.3
KernelSHAP	1	0.4	1.4
DiCE	2	0.0	**2.0**
OSDT	0	0.5	0.5

5.6 F5 - Interactivity

5.6.1 F5 : Normative definition

This property quantify the interactivity of the method. Ideally, explanations should allow the explainer and the explainee to interact with each other, e.g. simply to adapt the explanation to the level of expertise or simply explore the explanations.
We can divide it into two sub-properties :

Interaction level (F5.1)

This property evaluates the degree to which the XAI methods allow interactivity with the user.

$$m_{f5.1} = \begin{cases} 2 & \text{interactive control is provided by default} \\ 1 & \text{no default interaction but it can be implemented (e.g. via API)} \\ 0 & \text{no interaction is provided, and implementing it is complex} \end{cases} \tag{5.13}$$

Controllability (F5.2)

The property addresses how the explanation method can be controlled and if it has built-in enhancing mechanisms such as enabling user feedback.

$$m_{f5.2} = \begin{cases} 4 & \text{full user control with dynamic feedback improving explanations} \\ 3 & \text{partial control with significant user influence on explanations} \\ 2 & \text{limited control with pre-defined options for refinement} \\ 1 & \text{minimal control (e.g. visual exploration)} \\ 0 & \text{no control over explanations} \end{cases} \tag{5.14}$$

5.6.2 F5 : Implementation and Results

F5.1 Interaction level : scores

LIME (1) : No default built-in interaction but it can be implemented (e.g. API, Shiny)

KernelSHAP (1) : Same justification as LIME.

DiCE (1) : Same justification as LIME.

OSDT (1) : Same justification as LIME.

F5.2 Controllability : scores

LIME (2) : Besides visual exploration, explanations can be refined by choosing specific features or adjusting the number of features shown, offering limited control.

KernelSHAP (2) : Similar justification as LIME.

DiCE (2) : Provides limited controllability via desired target class, set of features allowed to vary, permitted value ranges, etc.

OSDT (1) : It does not provide any interactive functionality after training besides visual exploration.

Summary The table 5.5 shows similar results across the board due to the absence of interactivity as a central feature in chosen methods.

Table 5.5: F5 Interactivity scores

	Interaction Level (F5.1)	Controllability (F5.2)	Total
LIME	1	2	**3**
KernelSHAP	1	2	**3**
DiCE	1	2	**3**
OSDT	1	1	2

5.7 F6 - Fidelity

5.7.1 F6 : Normative definition

This property measures the extent to which the explanations is close to the true decision-making process of the underlying model.
For instance, it's crucial to know whether the information comes from a surrogate model, makes linearity assumptions about the model or neither of these.

Fidelity check (F6.1)

This property simply reflect the fidelity to the underlying model, i.e. "penalizing" techniques making linear-assumption or using surrogate model.
The goal is to highlight misleading conclusion on complex non-linear data draw by methods making unrealistic assumptions in this context, where explainability is the most important in this framework.
Therefore, we reward techniques without the aforementioned qualities.

$$m_{f6.1} = \begin{cases} 1 & \text{no surrogate model or linearity assumptions are used} \\ 0 & \text{a surrogate model OR a linearity assumptions are used} \end{cases} \tag{5.15}$$

Surrogate Agreement (F6.2)

This assesses, —if a surrogate model is used— the extent to which its predictions align with those of the black-box model.
The evaluation metric normalize the average prediction difference between the black-box and the surrogate model.

The solution is a score between 0 and 1 translating the alignment in prediction between the surrogate and black-box model.
When no surrogate model is used, the score is assign to the maximum value $m_{f6.2} = 1$

$$m_{f6.2} = 1 - \frac{\sum_{i=1}^{N} |b(x_i) - s(x_i)|}{N \max(b(x))} \tag{5.16}$$

where :

- $b(x_i)$ is the prediction of the black-box model for the instance x_i
- $s(x_i)$ is the prediction of the surrogate model for the instance x_i
- N is the number of instances used for the evaluation
- The denominator is simply the number of samples evaluated times the maximum value of the black-box model predictions. In classification cases $\max(b(x)) = 1$.

It should be mentioned that surrogate agreement evaluates faithfulness of the surrogate model to the black box, not the human interpretability or causal validity of the explanation. This dissociates F6 Fidelity metric from the F7 Faithfulness metric

5.7.2 F6 : Implementation and Results

F6.1 Fidelity check : scores

LIME (0) : It uses a surrogate model to locally approximate the behavior of the black-box model.

KernelSHAP (0) : It estimates Shapley values via a weighted linear regression fitted on simplified samples. It constitutes an approximation mechanism used to derive the explanation, is therefore considered surrogate-based.

DiCE (1) : It does neither rely on surrogate model nor on any linearity assumptions.

OSDT (1) : The method does neither make a linearity assumption nor use a surrogate model.

F6.2 Surrogate Agreement : scores

LIME (0.7) : Surrogate agreement is evaluated by comparing the predictions of the local linear surrogate learned by LIME with the predictions of the underlying neural network, following Equation 5.16. Agreement is computed over a set of randomly sampled instances, providing an empirical estimate of how well the surrogate reproduces the black-box model predictions beyond the single explained instance. The reported score corresponds to the mean surrogate agreement, with uncertainty quantified via a 95% confidence interval under an IID sampling assumption $[0.6617, 0.7383]$.

KernelSHAP (0.8) :

The same surrogate agreement procedure is applied to SHAP : predictions are reconstructed using the additive attribution model (base value plus feature attributions). Higher agreement is expected due to SHAP's theoretical guarantees of local accuracy, and the resulting score reflects the extent to which these guarantees hold empirically for the chosen dataset and target class. The 95% confidence interval is $[0.7653, 0.8347]$.

DiCE (1) : Does not use a surrogate model.

OSDT (1) : Does not use a surrogate model.

Summary The table 5.6 showcase the limited scope of the fidelity metric : DiCE and OSDT yield maximum result by not using surrogate or linearity assumption.

Table 5.6: F6 Fidelity scores

	Fidelity Check (F6.1)	Surrogate Agreement (F6.2)	Total
LIME	0	0.7	0.7
KernelSHAP	0	0.8	0.8
DiCE	1	1.0	**2.0**
OSDT	1	1.0	**2.0**

5.8 F7 - Faithfulness

5.8.1 F7 : Normative definition

In this property, we assesse how reliably the XAI method capture the behavior of the black-box model, whatever the scope. It's important to make the distinction between faithfulness and fidelity. Fidelity checks the alignment between the explanation model and the black-box model, in the case of post-hoc methods, and the whether the model makes linearity assumption. While Faithfulness, intents to compare the *explanations* given by the model with the nature of the black-box model.

Incremental Deletion(F7.1)

This property defines how the progressive removal of input features identified as relevant by the XAI methods impacts the predictive model's output f.

$$m_{f7.1} = \frac{1}{N}\sum_{j=1}^{N}\frac{\int_0^n \Big(f_j(i_r) - f_j(i)\Big)di}{\int_0^n f_j(i_r)di} \tag{5.17}$$

where :

- N is the number of instances evaluated
- n denotes the number of incrementally removed features.
- f_j is the black-box model prediction for the j^{th} instance".
- i_r represents the i^{th} feature removed based on a random explainer.
- i corresponds to the i^{th} relevant feature according to the XAI method.
- The integral $\int_0^n$ is the area under the curve (AUC) of the model's predicted probability after removing k features $(0 \leq k \leq n)$. A smaller AUC means the probability collapses quickly, indicating a correct importance ranking of the method, i.e. a faithful explanation.

RemOve And Retrain (ROAR) (F7.2)

This sub-property is tailored for global explanation methods. It compare the accuracy of the model after incrementally removing important features following the XAI methods and the accuracy of the model after removing random features.
To quantify, it estimates the area between the curves. (This metric was not computed for any ofthe studied methods because it requires global feature ranking, which is not avaibble for DiCE and OSDT, and is not meaningful in my

opinion for both LIME and SHAP. The global feature ranking provided by SHAP is more of a feature summary than a true global feature importance ranking.)

$$m_{f7.2} = \begin{cases} \frac{\int_0^n \big(a(i_r) - a(i)\big) di}{\int_0^n a(i_r) di} & \text{for classification task} \end{cases} \tag{5.18}$$

where :

- $a(i_r)$ is the accuracy of the model after retraining without a random feature.
- $a(i)$ is the accuracy of the model after retraining without a important feature following the XAI method.

White-box Check(F7.3)

The White-box Check's goal is again to test if the explanation model truly grasps the underlying reasoning of black-box model.
It uses a white-box surrogate model trained to mimic the black-box one and It will serve as explanation. His metric uses the proportion of sample for which both the surrogate model and the black-box model are aligned.
The surrogate model used depends on the explanation format. For instance, linear regression model serves effectively as white-box for LIME.
In this work, the white-box check will be measured on control synthetic data to test the explanation's model capacity to truly identify the model's reasoning.

$$m_{f7.3} = \begin{cases} 3 & \text{agreement} \geq 95\% \\ 2 & 80\% \leq \text{agreement} < 95\% \\ 1 & 60\% \leq \text{agreement} < 80\% \\ 0 & \text{agreement} < 60\% \end{cases} \tag{5.19}$$

5.8.2 F7 : Implementation and Results

F7.1 Incremental Deletion : scores

This metric was only evaluated for one class, "CAQ", due to lack of of time and of sufficient large amount of data points for minority class. The score for all the methods were computed on the same subset of individuals.
Instead of a Deletion, one decided to perturb the important features defined by the model in order to avoid having to retrain the model, which is sensitive to feature's order.
For DiCE and OSDT a concept as close as possible to feature importance was inferred, due to its absence in their design.

LIME (0.8) : Incremental deletion is computed by perturbing values in a decreasing order of importance defined by LIME instance-level feature's attribution. The probability decay curve is summarized via the area under the curve (AUC). A random-feature ordering is used as a baseline to normalize the score and control for arbitrary perturbations.

KernelSHAP (0.6) : Uses the same methodology as LIME with ordering the perturbation by SHAP attribution.

DiCE (0.5) : For DiCE, Incremental Deletion (F7.1) is adapted to the counterfactual explanation setting by interpreting feature importance through counterfactual feature changes. For each evaluated instance, counterfactuals are generated toward all alternative target classes. Features whose values differ from the original instance are treated as explanatory and prioritized in a feature ranking.
Faithfulness is assessed by incrementally reverting counterfactual features back to their original values and measuring the degradation of the predicted probability of the target class. The resulting probability curves

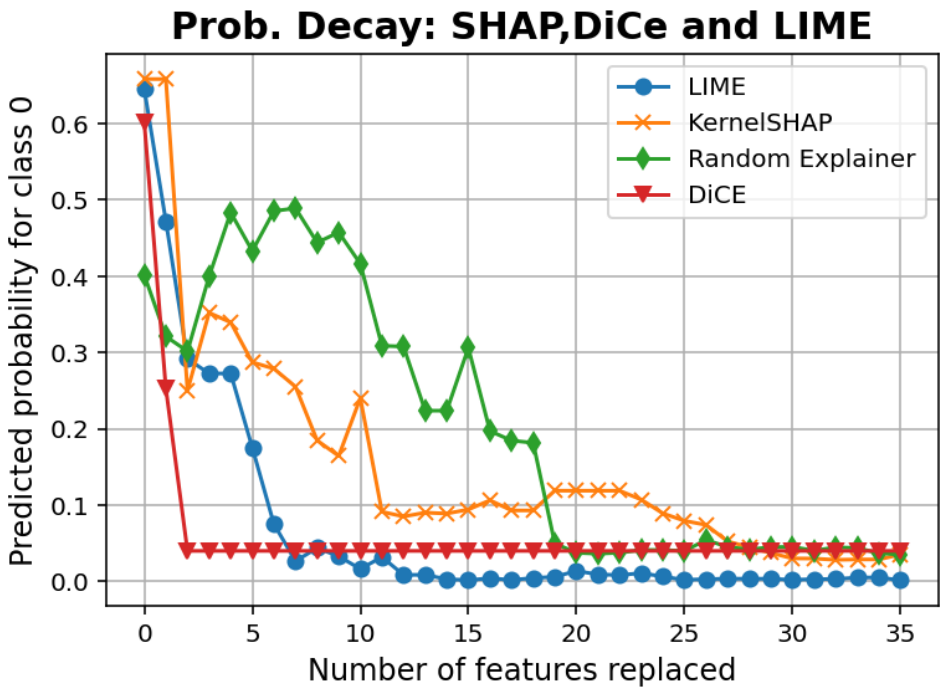
are summarized using the area under the curve (AUC). A random-feature baseline is used to control for arbitrary perturbations. Scores are averaged across counterfactuals, target classes, and instances to obtain the final F7.1 score.
Higher scores indicate that the features modified by DiCE are functionally responsible for the counterfactual prediction, providing evidence that the generated explanations faithfully reflect the model's decision logic.

OSDT (0.6) : Here "important features" for an explanation are the features on its decision path. The importance is structural. We rank them with the idea : the earlier a feature is on the decision, the more important it is, because an alternative decision would lead to drastically different decision path.
OSDT operates on a binary features sub space resulting from thresholding and categorical encoding. The incremental deletion becomes flipping the important binary features. It becomes evaluating label stability rather than probability decay, as the model is deterministic.
The deletion becomes a gradual flipping, starting from the most important to the least, of the binary features along the input's decision path. A binary indicator records whether the prediction changes with the modification of on (decision) path feature k :

$$\mathbb{I}\big[\hat{y}(x^k_{on}) \neq \hat{y}(x)\big] \tag{5.20}$$

For each k feature, we randomly change k off-path features multiple times, to try to reduce the variability of randomly choosing a feature. Then we average the binary indicators vector quantifying whether the prediction changes or not, to a scalar representing the average change in the input prediction due to changing an off-path feature. This quantity, by design, is always zero : flipping the value of feature that's not on the decision path will not impact the prediction.
Each instance yields a deletion curve describing how prediction stability degrades as a growing fraction of explanation-relevant features is removed.
To summarize this behavior independently of explanation length, deletion steps are expressed as a fraction of the decision path, and curves are aggregated across instances. Faithfulness is then quantified by the area under the resulting degradation curve, which captures both the speed and magnitude of prediction collapse. The computation becomes $F_{7.1} = auc_{\text{on path}} - auc_{\text{off path}}$, the latter being 0.

The following plot shows the probability decay induced by Incremental Deletion for KernelSHAP, LIME and DiCE for a single instance. The plot clearly shows the correct identification of important features by the method by emphasizing a difference between method's feature importance and random feature selection.



F7.2 ROAR : scores

The ROAR metric was ommitted from the analysis because it requires a global feature ranking, which LIME and SHAP are unable to meaningfully provide. DiCE is a instance-level explanation and does not fournish any informations about global feature importance. Although global feature information could be derived from OSDT, the time required to implement it and the unavailability of any possible comparison with other metrics lead to an non implementation of it.

F7.3 White box check : scores

LIME (1) : A synthetic classification task with a known linear data-generating process is created. Ground-truth feature influence, i.e. instance level-deviation from a baseline prediction, is computed analytically and serves as a reference explanation. LIME's explanations are directly compared to this influence vector. The evaluation assesses whether the method recovers the true feature influence structure under perfect conditions. LIME reached a agreement of 62%.

KernelSHAP (2) : Same explanation as LIME. KernelSHAP exhibits the highest agreement, 91% among the post-hoc metrics.

DiCE (0) : Same setup as LIME and KernelSHAP, but DiCE, being a example-based method, is evaluated through relevance inferred from counterfactual feature changes. A feature is marked as relevant if it changes in at least half of the generated counterfactuals. Agreement is computed as the proportion of features for which DiCE's relevance decision matches the ground truth relevance decision. The latter being defined as a binary indicator labelling the most relevant features on the instance level. DiCE agreement is 53%.

OSDT (2) : A synthetic classification task is constructed using an simple rule-based decision tree, which serves as the true white-box mechanism and ground truth. OSDT is trained on a binarized predicate representation, as it is required. For each test instances, the model's prediction is compared to the ground-truth rule output. Faithfulness is simply computed as the proportion of instances for which both decisions agree, i.e. the model correctly identify the reasoning of the underlying problem. OSDT yielded an agreement of 92%

One understands that this test is trivial and does not prove much besides the model's capabilities in simple framework.

Summary The table 5.7 highlights the two most faithful methods as defined in this framework : OSDT and KernelSHAP. DiCE's score reflects the difficulty in adapting example-based methods to certain criteria of the benchmark, which were mainly designed for post-hoc attribution methods.

Table 5.7: F7 Faithfulness scores

	Incremental Deletion (F7.1)	ROAR (F7.2)	White-Box Check (F7.3)	Total
LIME	0.8	NaN	1	1.8
KernelSHAP	0.6	NaN	2	**2.6**
DiCE	0.5	NaN	0	0.5
OSDT	0.6	NaN	2	**2.6**

5.9 F8 - Truthfulness

5.9.1 F8 : Normative definition

This property checks if the explanations are aligned with common knowledge of the user's true world. It includes being accordant with prior relevant domain knowledge and beliefs of the "explainee" but also to detect biased models.

Reality Check (F8.1)

This property test if the XAI methods prevents the generation of unrealistec data samples, ensuring alignment with real-world knowledge. For instance, a feature age should not be negative.
It is divised into two sub-properties :
Feature constraints consistency ($m_{f8.1A}$) :It ensures that the generated explanations do not violate feature bounds.
Feature correlation consistency ($m_{f8.1B}$) : It ensures that the generated explanations follows the same observed correlation as in the training data.
The final score range from 0 (least realistic) to 2 (most realistic)

$$m_{f8.1} = m_{f8.1A} + m_{f8.1B} \tag{5.21}$$

$$m_{f1.3A} = m_{f8.1A} = \begin{cases} 1 & \text{expl. respect feature constraints} \\ 0 & \text{they does not} \end{cases} \qquad m_{f1.3B} = m_{f8.1B} = \begin{cases} 1 & \text{expl. represt feature correlations} \\ 0 & \text{they does not} \end{cases} \tag{5.22}$$

Bias Detection (F8.2)

This property evaluates whether the XAI method can reveal biases within the model or dataset. By accomplishing this property, the method goes beyond the scope of simply giving information, overall, it help improving the model reliability.
This method can be implemented using biased models or synthetic data to showcase the XAI method's ability to expose biases. It includes Husky-vs-Wolf Classifier,Gendered Occupation Models, Simulated Bias (through

Synthetic Data), which will neither be used nor study in this work.
In this work, Bias Detection measures potential to reveal bias, not guaranteed bias identification.

$$m_{f8.2} = \begin{cases} 1 & \text{bias is exposed by the XAI method} \\ 0 & \text{bias is not detected} \end{cases} \tag{5.23}$$

5.9.2 F8 : Implementation and Results

F8.1 Reality Check : scores

LIME (0 + 0) : A) LIME does not enforce hard constraints on feature validity when doing local perturbations. Nonetheless, perturbations are made in the vicinity of an existing point. If allowed, the score would be of 0.5. B) It perturbs features independently when generaing local samples, without enforcing any joint feature dependencies.

KernelSHAP (0 + 0) : A) It relies on background data to approximate feature contributions through sampling. This strategy uses empirical data distribution, and de facto following the world-reality of the data. Nonetheless, it does not enforce explicit feature constraints. As of LIME, the score would be of 0.5. B) KernelSHAP assumes feature independence when estimating Shapley values through sampling.

DiCE (1 + 0) : A) DiCE enable user-defined feature constraints such as permitted ranges. The contraints are enforced during CF generation, ensuring all explanations remain consistents. B) DiCE does not explicitly model or enforce feature correlations during counterfactual generation. It does not guarantee that generated CFs preserve obsersed dependencies in the training data.

OSDT (1 + 1) : A) The method does not generate synthetic instances as part of explanation, therefore it cannot violate feature constraints via generation.
B) Correlation consistency is defined over generated explanation instances. Since OSDT produces no synthetic instances, the metric is non-diagnostic for this method and is satisfied by default.

F8.2 Bias Detection : scores

Following the definition, all the studied methods have the tools be able to reveal bias by studying their outputs. For instance, LIME and KernelSHAP can highlights systematic dominance of certain features.

Summary The table 5.8 shows an outperforming OSDT method, but it is mostly dued to the lack of the metric's capabilites and scope for such model.

Table 5.8: F8 Truthfulness scores

	Reality Check (F8.1)	Bias Detection (F8.2)	Total
LIME	0	1	1
KernelSHAP	0	1	1
DiCE	1	1	2
OSDT	2	1	**3**

5.10 F9 - Stability

5.10.1 F9 : Normative definition

This property ensure that XAI method is robust to small changes in the input. It is divided into two sub-property.

Similarity (F9.1)

The property checks whether similar points (neighbors), belonging to the same class, have similar explanations. This method for defining neighbors can vary based on the task and data type.

$$m_{f9.1} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{n}\sum_{j=1}^{n}\frac{1}{1+d(e_i,e_i^{(j)})} & \text{if pairwise distance is meaningful}\\ \\ 1-\frac{1}{k}\sum_{l=1}^{k}\sigma_l/\mu_l & \text{otherwise}\end{cases} \tag{5.24}$$

Where :

- N is the number of evaluated instances
- n corresponds to the number of neighboring samples for each evaluated instance.
- e_i is the explanation generated for the reference instance i^{th} .
- $e_i^{(j)}$ is the explanation generated for the neighbor j or reference instance i.
- $d(e_i, e_i^{(j)})$ measure the pairwise distance between the reference explanations e_i and it's neighbors $e_i^{(j)}$
- k corresponds to the number of components in the explanation
- σ_l is the standard deviation of the l^{th} component across neighbors
- μ_l represent the mean of the l^{th} component across neighbors.

Identity (F9.2)

This property mesure the variability of the explanation method. For an identical input, rather than similar ones like in F9.1, the method is use multiple times to measure its consistency. A higher variability indicates greater instability in the XAI method. Identity metric isolates intrinsic stochasticity in the explanation algorithm itself.

$$m_{f9.2} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{r}\sum_{h=1}^{r}\frac{1}{1+d(e_i^{(0)},e_i^{(h)})} & \text{if pairwise dist. is meaningful}\\ \\ 1-\frac{1}{k}\sum_{l=1}^{k}\sigma_l/\mu_l & \text{otherwise}\end{cases} \tag{5.25}$$

Where :

- N denotes the number of evaluated instances
- r represent the number of repeated runs for the same instance
- $e_i^{(h)}$ represents the explanation generated for the instance i at the h^{th} run.
- $e_i^{(0)}$ represents the reference explanation for the instance i.
- k corresponds to the number of components in the explanation
- σ_l is the standard deviation of the l^{th} component across runs
- μ_l represent the mean of the l^{th} component across runs.

5.10.2 F9 : Implementation and Results

The same underlying idea was used for all methods : A points are randomly sampled, called anchor points; and N nearest point based on the Gower distance were selected. The Gower distance is distance metric deisigned for datasets containing heterogeneous feature types. It defines the distance between two instances as the average of feature-wise dissimilarities, where each feature contributes a value in $[0, 1]$, ensuring comparability across scales and data types. In this work, feature-wise dissimilarties are computed as follows :

- Categorical features contribute 0 if values are equal and 1 otherwise.
- Ordinal features contribute the absolute difference between values, normalized by the feature's range.
- Continuous features contribute the absolute difference between values, normalized by the observed range.

The final distance is obtained by averaging these contributions across all fetures.
The points all belong to the same classes, "CAQ". It overestimates the true similarity across the full data set distribution, but is acceptable since all methods use the same standard.

F9.1 Similarity : scores

LIME (0.1), KernelSHAP (0.1) : Explanation similarity is evaluated by measuring the proximity of attribution or surrogate coefficient vectors across A neighborhoods containing N similar instances.
After a feature-wise normalization, similarity is computed between a reference instance and its neighbors using a distance-based similarity defined as follows :

$$\text{sim}_a = \frac{1}{N-1} \sum_{j=2}^{N} sim(e_1, e_j) \tag{5.26}$$

Where

$$e_i = \begin{cases} \frac{\phi(x_i)-\mu_a}{\sigma_a+\epsilon} & \text{For KernelSHAP,} \\ \\ \frac{\beta(x_i)-\mu_a}{\sigma_a+\epsilon} & \text{For LIME} \end{cases} \tag{5.27}$$

and μ_a is mean vector for all the features $f \in \mathcal{F}$ in batch a and σ_a is the standard deviation vectors for all features $f \in \mathcal{F}$ in the batch a and $\text{sim}(e_1, e_j) = \frac{1}{1+\|e_1-e_j\|_2}$. The similarity is averaged locally and then globally. This metric captures the stability of attribution patterns (KernelSHAP), and local surrogate models (LIME) under small input perturbations.

DiCE (0.4) : The similarity notion is not naturally defined for example-based explanations. For each batch a of similar instances, counterfactual explanations are generated independently. Since DiCE explicitly promotes diversity among genreated counterfactuals, a canonical counterfactual, i.e. the closest following Gower distance of the anchor point, per instance, is selected. Then, each canonical CF is represented by a binary feature-change vector $S_i = \{f \in \mathcal{F} | x_{i,f}^{cf} \neq x_{i,f}\}$ where f is a feature of the feature set $\mathcal{F} = \{1, ..., d\}$, indicating which features differ between the original instance i and its canonical counterfactual cf.

This representation captures which features must be modified, rather than the magnitude of those modifications, unquantifiable under DiCE. Similarity within neighborhood a is computed as the mean pairwise Jaccard similarity between the feature-change vectors. It measures the extent to which nearby instances require changing the same features to obtain the same class, Formally, for batch a of N instances, similarity is defined as :

$$\text{sim}_a = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{5.28}$$

Where S_i the feature change-vector of instance i, S_j is the feature change-vector of instance j, we sum over $i < j$ to not take into consideration the same pairs, $\frac{2}{N(N-1)} = frac1\binom{N}{2}$ is the total number of terms. This

quantity measures the extent to which nearby instances require modifying the same features to obtain the same target class.

OSDT (0.8) : For the OSDT metric, the similarity of explanations is the similarity of decision paths. The distance between the original point's explanation, i.e. decision path, and every points of the batch is computed. The distance is $1 - L/M$, where L is the number of common split (same feature, and same boolean evaluation), between the two decision path and M is the maximum length decision path between the two. Explanations not sharing any common split are assigned a maximum distance of 1.

The similarty is computed across batch as follow and then average :

$$\text{sim}_a = \sum_{n=1}^{N} \mathbb{E}\Big(\frac{1}{d_n}\Big) \tag{5.29}$$

Where d_n is the distance between the anchor point's decision path and the n of the same batch a ,$a = 1, ..., A$ and is the number of batch, and $n = 1, ..., N - 1$ is the number of points per batch.
The final similarity is the average of similarities across batch.

F9.2 Identity : scores

LIME (0.1), KernelSHAP (0.2) : For KernelSHAP and LIME, which both rely on stochastic sampling procedures, repeated explanations are expected to exhibit variability. F9.2 quantifies this variability by measuring the dispersion of attribution or surrogate coefficient vectors across runs, thereby isolating intrinsic randomness from input-induced instability. The same data points were used as the A anchors points as in the previous section. Explanations are generated 10 times per instances.

DiCE (0.9) : Identity is assessed in terms of consistency of the counterfactual explanations generated for the same input across R runs. For this work, the identity metric is sub-divided into two sub-metrics.

The first notion is called feature-identity and focuses on feature-level consistency : whether DiCE consistenly identifies the sames features are requiring modification to change the prediciton. It executes the same steps as in the previous section, i.e. canonical CF into feature-change vector (see section 5.10.2). After the computation of the R feature-change vector, the metric is computed as the mean pairwise Jaccard similarty between these feature-change vectors :

$$\text{Id}_{\text{feature}}(x) = \frac{2}{R(R-1)} \sum_{1 \le k < \ell \le R} \frac{|S_k(x) \cap S_\ell(x)|}{|S_k(x) \cup S_\ell(x)|}. \tag{5.30}$$

Where S_k the feature change-vector of instance k, S_ℓ is the feature change-vector of instance ℓ, we sum over $i < j$ to not take into consideration the same pairs.
A high feature-identity score indicates that DiCE consistently modifies the same subset of features across repeated runs for a given instance, suggesting stable identification of relevant features.
The second metric is the distance-identity : it quantifoes how close, within a set of R repetitions, for one instance i, are its canonical CF. The pairwise distances aer computed using the Gower distance. The distance-identity score for given instance is obtained :

$$\bar{d}(x_i) = \frac{2}{R(R-1)} \sum_{1 \le k < \ell \le R} d_G\big(x_i^{cf,k}, x_i^{cf,\ell}\big) \tag{5.31}$$

Where $x_i^{cf,k}$ is k counterfactual cf of the instance i.

The mean distance $\bar{d}(x_i)$ is converted into a similarity-based score via :

$$\text{Id}_{\text{distance}}(x_i) = \frac{1}{1 + \bar{d}(x_i)} \tag{5.32}$$

The global score is obtained by averaging across all instances for which at least two canonical counterfactuals are availables. The final score for DiCE is :

$$\text{F9.2}_{\text{DiCE}} = \frac{1}{2}\text{Id}_{\text{feature}} + \frac{1}{2}\text{Id}_{\text{distance}} \tag{5.33}$$

This balanced aggregation reflects the dual nature of counterfactual explanations, which must be stable both in feature selection and in proposed actions.

OSDT (1) : For OSDT, explanations are deterministic and correspond exactly to the model's decision path. Therefore, the identity metric F9.2 is structurally maximized (equal to 1) for all instances, and does not provide additional discriminative power.

We report the maximum value by construction.

Summary In table 5.9, we can observe drastically different performance along methods. KernelSHAP and LIME, as expected, perform poorly due to its stochastic nature for one and its scope for the other. DiCE, despite it is stochastic nature, perform surprisingly well. The implementation choices might be a credible explanation for the unexpected success : they are implemented in a way to curb its diversity property and selecting only the closest CF generated. OSDT is most stable method under this framework, which can explain its deterministic nature.

Table 5.9: F9 Stability scores

	Similarity (F9.1)	Identity (F9.2)	Total
LIME	0.1	0.2	0.3
KernelSHAP	0.1	0.2	0.3
DiCE	0.4	0.9	1.3
OSDT	0.8	1.0	**1.8**

5.11 F10 - (Un)Certainty

5.11.1 F10 : Normative definition

This property provide the level of confidence in the XAI method's output to end-users, i.e. non ML-practitioners. It's score from 0 (no confidence measures) to 5 (fully transparent confidence reporting).

$$m_{f10} = \sum_{i=1}^{5} c_i \tag{5.34}$$

Where c_i corresponds to a binary indicator for the i^{th} confidence aspect :

- $c_1 = 1$: Confidence in the black-box model's result is reported (e.g. by displaying the model accuracy)
- $c_2 = 1$: Confidence in the XAI explanation is reported.
- $c_3 = 1$: Random processes in explanation generation are disclosed (e.g., sampling/random perturbation)

- $c_4 = 1$: The instance's distance from the training data distribution is indicated.
- $c_5 = 1$: Addition confidence indicators/measures are provided.

5.11.2 F10 : Implementation and Results

LIME (2) :

C1) The method report confidence in the black-box model's out. It display the probability of the outcome. Score of 1. *C2)* Confidence in the explanation not reported. *C3)* Random processes are not disclosed. *C4)* Distance from the training data distribution not indicated. *C5)* LIME provides de R^2 scores, the coefficient of determination. It represents the proportion of the variation in the dependent variable that is predictable from the independent variables.

KernelSHAP (1) :

C1) The method report confidence in the black-box model's out. It display the probability of the outcome. Score of 1. *C2)* Confidence in the explanation not reported. *C3)* Random processes are not disclosed. *C4)* Distance from the training data distribution not indicated. *C5)* No substantial additional confidence indicators.

DiCE (2) :

C1) It shows the explanation outputs. Score of 1. *C2)* Confidence in the explanation not reported. *C3)* Not openly disclosed. *C4)* Distance from the training data distribution not indicated, but distance from the decision boundary can be deduced. Score of 1. *C5)* No substantial additional confidence indicators.

OSDT (1) : *C1)* The explanation output does not explicitly report confidence measures (e.g. probabilty) associated with the predicted outcome. *C2)* While OSDT produces an optimality certificate under its theoretical framework, it does not provide user-facing measures quantifying the confidence of individuals explanations. *C3)* No random processes to be disclosed. Score of 1. *C4)* The instance's distance from the training data distribution is not indicated. *C5)* No substantial additional confidence indicators.

Summary The table 5.10 shows that DiCE and LIME project the highest level of confidence toward non-ML practitioners.

Table 5.10: F10 (Un)Certainty scores

	Total
LIME	**2**
KernelSHAP	1
DiCE	**2**
OSDT	1

5.12 F11 - Speed

5.12.1 F11 : Normative definition

This property assesses the computation time required by the XAI method to generate and explanation. Its purpose is to quantify the real-world readiness of the methods. The computation time is computed from initialization to the production of the first explanation.

$$m_{f11} = \begin{cases} 4 & t \leq 0.1\text{sec} \\ 3 & 0.1 < t \leq 1\text{sec} \\ 2 & 1 < t \leq 5\text{sec} \\ 1 & 5 < t \leq 10\text{sec} \\ 0 & t > 10\text{sec} \end{cases} \tag{5.35}$$

5.12.2 F11 : Implementation and Results

Summary The table 5.11 shows that LIME is quickest studied explainer, followed by DiCE. As expected KernelSHAP is the least performing post-hoc methods, due to its expensive approximation strategy and its guaranteed axioms. And the framework is not fitted for intrinsic interpretable model : we are comparing model training and explanation generation versus solely explanation generation.

Table 5.11: F11 Speed scores

	Total
LIME	**3**
KernelSHAP	1
DiCE	2
OSDT	0

Chapter 6

Results and Conclusion

6.1 Results

As already mentioned in the beginning of section 5, the results are to be understand as what they are : a comparison between XAI methods belonging to different model's families, under a benchmark framework primarly thought for post-hoc attribution based methods, in the context of this classification task on (mostly) categorical data and not-so-accurate predictions models. Some methods will be advantaged or disadvantaged depending on their nature and the context of the evaluation.
Therefore, all results should be interpreted with caution, not as ground truth, and takin in the aforementioned context.For comparison purpose, all the metric have been normalized using min-max scaling.

The benchmark reaveals, distinct, almost, non-overlapping performances profiles across methods, with no method dominating across all functional properties, figure 6.1 illustrates it.

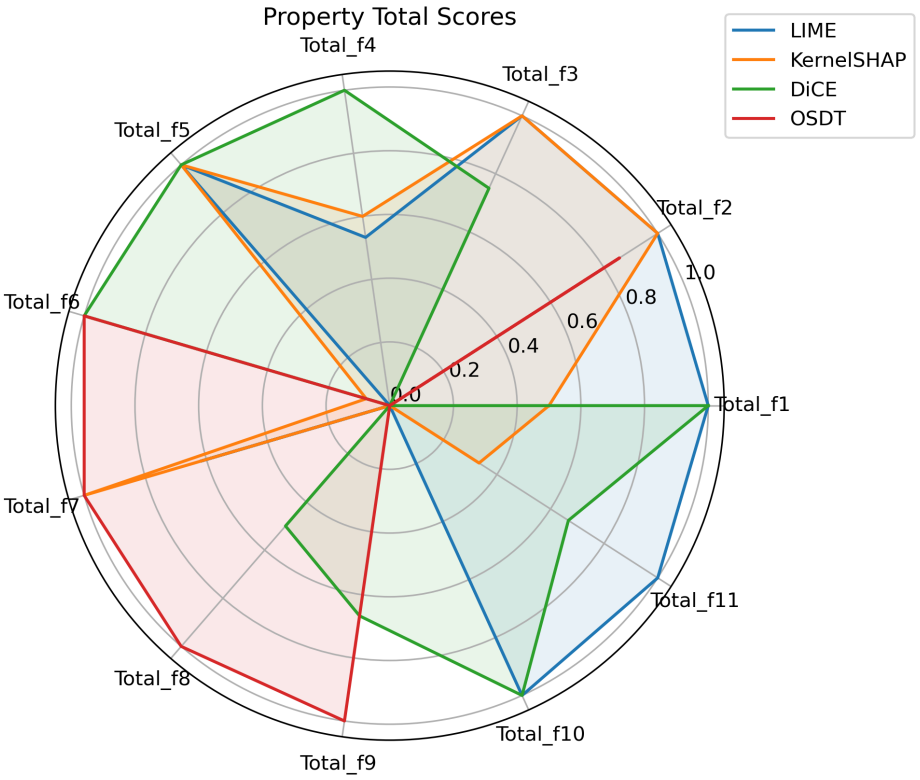


Figure 6.1

The two plots 6.2 shows the total scores for SHAP and LIME on the left and DiCE and OSDT on the right. It perfectly illustrates the domain of expertise of certain methods : LIME and KernelSHAP perform well on the Representativeness(F1), Selectivity(F3), Structure (F2), Speed (F11) and Certainty (F10) metrics, which can be labelled as "Communication-oriented properties" ; OSDT performs well on "Model-alignement" properties : the Fidelity (F6), Faithfulness (F7), Truthfulness (F8) and Stability (F9) ; and DiCE on constrastive reasoning (F4).
The benchmark underlines intuitives knowledge linked to the methods' structure : OSDT is more stable than LIME and SHAP ; DiCE gives the more contrastive information. It is a sign that the benchmark is able to effectively captures the strengths and weaknesess of methods in its current shape.

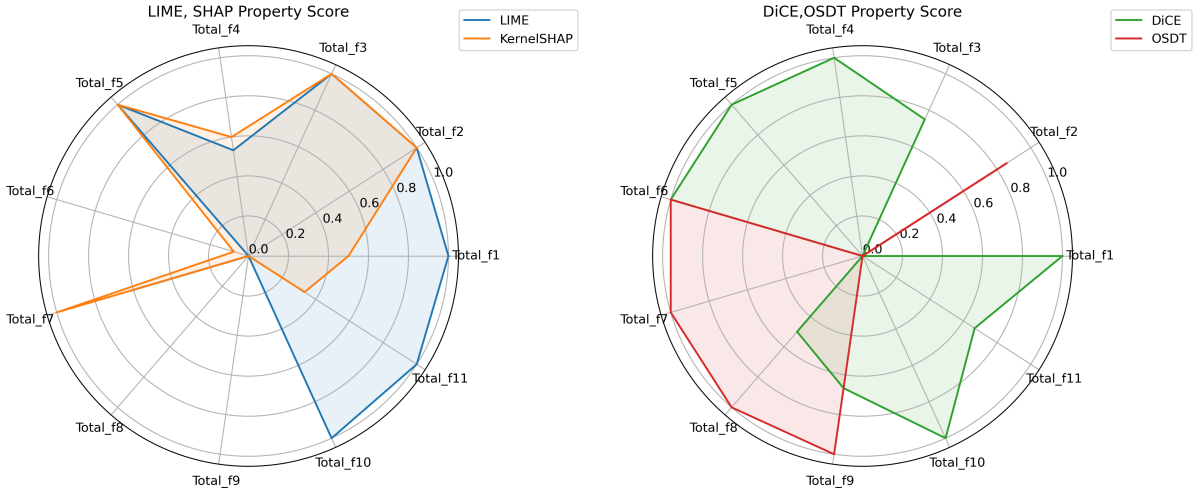


Figure 6.2: Comparison of the predictions and intentions of voters

6.2 Discussion

This work applied a functionally ground benchmark to compare four XAI approaches that produces different explanatory objects (attributions scores, counterfactual instances, and decision-path logic). In this setting — multi-class prediciton on an almost fully categorical tabular dataset— the benchmark does not induce a single ranking of methods. Instead, it reveals distinct performance profiles, aligned with different functional goals and architecture, supporting the view that interpretability is a multi-faceted rather than reducible to a single scalar score.

The Stability property (F9) effectively separates methods with intrinsic stochasticity (or locality dependence) from deterministic explanation mechanisms. The Identity and Similarity sub-metrics produce a clear ranking from attribution-based methods (LIME/KernelShap) to intrinsic interpretable model (OSDT), consistent with the fact that local surrogate fitting and sampling-based approximations introduce variability across runs. In contrast, OSDT explanations are uniquely determined by the learned decision structure and the input instance.
Hower, DiCE's second place should be interpreted in the light of the implementation choice to select a canonical "closest counterfactual", which curb its diversity objective.
Furthermore, the Constrastivity property (F4), as mentioned in the previous section, behaves as intented in the sense that it isolates the explanatory paradigm that explicitly encodes alternatives : DiCE.

Despite being presented as general-purpose for post-hoc methods, several properties appears to have been formulated with post-hoc attribution outputs in mind, and this assumption becomes visible when applying them example-baed explanations. For instance, F2 (Structure) rewards graphical integrity via signed positive/negative contributions.

Interactivity (F5) appears to lack discriminability with is near-constant values across methods in this study. It is largely because the scoring captures whether some form of parameter control and *potential* interface implementation (e.g. API) are present.

Finally, several methodological constraints limit the scope of the numerical comparisons. Faithfulness (F7) and stability (F9) were evaluated only for one class due to both time and sample size constraints. ROAR (F7.2) was not computed in this setting, reducing the reach of F7's claims.
Parts of the evaluation rely on *paradigm translations* (e.g. mapping counterfactual changes to feature importances proxies), which are necessare for cross-method comparison but introduce additional assumptions.
Furthermore, the uncertainty introduced by the modest accuracy of the models, mitigated by choice to analyze points in which the model was highly confident of, adds another layer of cautions with the results of this analysis.
The models were trained on a single training-testing split. The perfect modus operandi would have been to add one black-box, one white box models ; train the models ; effectuate my whole analysis, for every class, for different training-test splits; repeat several times on each splits to remove the bias introduced but the splitting choice.

Interpreted through the AI HLEG [20] transparency framing (Explainability, Traceability, COmmunication), the benchmark results translate into practical selection guidance rather than a global ranking of methods. In this case study, communication-oriented properties — captured by F2 or F3— favor attribution-based methods or example-based methods. Whereas explainability, as model alignement — capturing Fidelity/Faithfulness/stability-type criteria — favors OSDT and decision path logic. Traceability is only indirectly implemented by the benchmark metrics and is instead supported here by the exploration of of the data and preprocessing transparency developed in earlier chapters.
The results do not revolutionaze our understanding, but instead offers guidance via a *concret* case analysis of the benchmark that aligns with real-world expectations.

6.3 Conclusion

This thesis compared *LIME*, *KernelSHAP*, *DiCE*, and an intrinsic interpretable model (*OSDT via GOSDT* solver) under a functionally grounded benchmark in a multi-class classification setting on an almost fully categorical tabular dataset. The evaluation does not yield a universally dominant method. Rather than being inconclusive, this outcome provides evidence that interpretability properties are *trade-off driven*, *dependent on the explanatory paradigm*—i.e., on whether explanations are expressed as feature contributions, counterfactual instances, or model-intrinsic decision logic. **Consequently, the benchmark's result supports goal-conditioned method selection rather than global ranking.**
Returning to the motivating application—predicting political-party alignment from lifestyle and demographic survey features—the results suggest that methods producing concise, communicable, instance-level summaries are best aligned with communication toward non-expert users. In this benchmark instantiation, attribution-based explanations (LIME/KernelSHAP) score higher on communication-facing properties such as representational structure and tunable selectivity, while counterfactual explanations (DiCE) provide the most explicit contrastive information by articulating how an outcome could change. In contrast, intrinsic explanations (OSDT) are most aligned with activity requiring high stability, exhibiting strong performance on model-alignment and reliability-related properties, including stability and faithfulness proxies.
More broadly, the benchmark does not collapse heterogeneous explanation types into a single ordering; it separates them into distinct functional profiles that correspond to different transparency goals. Taken together, the results support interpretability as a multi-faceted construct: the studied methods do not compete along a single latent "interpretability" axis, but instead realize different functional goals—*communication*, *contrastive reasoning/recourse*, and *model-aligned transparency*. This framing reconciles the absence of a single winner with the

fact that each method remains valuable under the goal it is designed to serve.
Finally, several limitations bound the strength and generality of the numerical comparisons. Some benchmark properties are weakly discriminative in practice (e.g., interactivity) or become partially non-diagnostic for certain paradigms, and cross-family method comparison sometimes requires semantic reinterpretation rather than mechanical metric reuse.

Bibliography

[1] "(PDF) Comprehensive Review of Deep Reinforcement Learning Methods and Applications in Economics". In: *ResearchGate* (Dec. 9, 2024). DOI: 10.3390/math8101640. URL: https://www.researchgate.net/publication/344351321_Comprehensive_Review_of_Deep_Reinforcement_Learning_Methods_and_Applications_in_Economics (visited on 02/06/2025).

[2] *2. Over-sampling — Version 0.13.0*. URL: https://imbalanced-learn.org/stable/over_sampling.html#naive-random-over-sampling (visited on 07/19/2025).

[3] *2. Over-sampling — Version 0.15.Dev0*. URL: https://imbalanced-learn.org/dev/over_sampling.html (visited on 12/13/2025).

[4] *A Liberal Plan for an Assertive, United, and Prosperous Quebec*. URL: https://plq.org/en/press-release/a-liberal-plan-for-an-assertive-united-and-prosperous-quebec/ (visited on 10/08/2025).

[5] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052. URL: https://ieeexplore.ieee.org/document/8466590 (visited on 02/21/2025).

[6] Sajid Ali et al. "Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence". In: *Information Fusion* 99 (Nov. 1, 2023), p. 101805. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101805. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001148 (visited on 02/21/2025).

[7] Elaine Angelino et al. *Learning Certifiably Optimal Rule Lists for Categorical Data*. Aug. 3, 2018. DOI: 10.48550/arXiv.1704.01701. arXiv: 1704.01701 [stat]. URL: http://arxiv.org/abs/1704.01701 (visited on 11/04/2025). Pre-published.

[8] Elaine Angelino et al. "Learning Certifiably Optimal Rule Lists for Categorical Data". In: ().

[9] *Artificial Intelligence*. In: *Wikipedia*. Feb. 2, 2025. URL: https://en.wikipedia.org/w/index.php?title=Artificial_intelligence&oldid=1273565751 (visited on 02/06/2025).

[10] *Artificial Intelligence and Digitalisation of Judicial Cooperation*. URL: https://eucrim.eu/articles/artificial-intelligence-and-digitalisation-of-judicial-cooperation/ (visited on 02/06/2025).

[11] Andrew Bell et al. "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 20, 2022, pp. 248–266. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533090. URL: https://dl.acm.org/doi/10.1145/3531146.3533090 (visited on 12/12/2025).

[12] Matthieu Bellucci et al. "Towards a Terminology for a Fully Contextualized XAI". In: *Procedia Computer Science* 192 (2021), pp. 241–250. ISSN: 18770509. DOI: 10.1016/j.procs.2021.08.025. URL: https://linkinghub.elsevier.com/retrieve/pii/S187705092101512X (visited on 12/11/2025).

[13] Leo Breiman. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)”. In: *Statistical Science* 16.3 (Aug. 2001), pp. 199–231. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1009213726. URL: https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full (visited on 07/09/2025).

[14] Dulce Canha et al. “A Functionally-Grounded Benchmark Framework for XAI Methods: Insights and Foundations from a Systematic Literature Review”. In: *ACM Comput. Surv.* 57.12 (July 14, 2025), 320:1–320:40. ISSN: 0360-0300. DOI: 10.1145/3737445. URL: https://dl.acm.org/doi/10.1145/3737445 (visited on 12/11/2025).

[15] Chaofan Chen et al. “This Looks Like That: Deep Learning for Interpretable Image Recognition”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html (visited on 12/12/2025).

[16] Eunsuk Chong, Chulwoo Han, and Frank C. Park. “Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies”. In: *Expert Systems with Applications* 83 (Oct. 15, 2017), pp. 187–205. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2017.04.030. URL: https://www.sciencedirect.com/science/article/pii/S0957417417302750 (visited on 02/06/2025).

[17] Evangelia Christodoulou et al. “A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models”. In: *Journal of clinical epidemiology* 110 (2019), pp. 12–22. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2019.02.004.

[18] *Coalition Avenir Québec*. In: *Wikipedia*. Oct. 7, 2025. URL: https://en.wikipedia.org/w/index.php?title=Coalition_Avenir_Qu%C3%A9bec&oldid=1315543357 (visited on 10/08/2025).

[19] *CORELS: Learning Certifiably Optimal RulE ListS*. URL: https://corels.cs.ubc.ca/corels/ (visited on 11/04/2025).

[20] *Ethics Guidelines for Trustworthy AI | Shaping Europe’s Digital Future*. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on 12/12/2025).

[21] Petko Georgiev et al. “Low-Resource Multi-task Audio Sensing for Mobile and Embedded Devices via Shared Deep Neural Network Representations”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.3 (Sept. 11, 2017), 50:1–50:19. DOI: 10.1145/3131895. URL: https://doi.org/10.1145/3131895 (visited on 02/06/2025).

[22] Sofie Goethals. “The Non-Linear Nature of the Cost of Comprehensibility”. In: (2022).

[23] Micah Goldblum et al. *The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning*. June 7, 2024. DOI: 10.48550/arXiv.2304.05366. arXiv: 2304.05366 [cs]. URL: http://arxiv.org/abs/2304.05366 (visited on 07/09/2025). Pre-published.

[24] Michael Greenacre and Raul Primicerio. *Multivariate Analysis of Ecological Data*. Bilbao: Fundación BBVA, 2013. 331 pp. ISBN: 978-84-92937-50-9.

[25] Robert C. Holte. “Very Simple Classification Rules Perform Well on Most Commonly Used Datasets”. In: *Machine Learning* 11.1 (Apr. 1, 1993), pp. 63–90. ISSN: 1573-0565. DOI: 10.1023/A:1022631118932. URL: https://doi.org/10.1023/A:1022631118932 (visited on 07/09/2025).

[26] Xiyang Hu, Cynthia Rudin, and Margo Seltzer. *Optimal Sparse Decision Trees*. Sept. 26, 2023. DOI: 10.48550/arXiv.1904.12847. arXiv: 1904.12847 [cs]. URL: http://arxiv.org/abs/1904.12847 (visited on 11/17/2025). Pre-published.

[27] *Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/ (visited on 09/10/2025).

[28] Ulf Johansson et al. “Trade-Off Between Accuracy and Interpretability for Predictive In Silico Modeling”. In: *Future medicinal chemistry* 3 (Apr. 1, 2011), pp. 647–63. DOI: 10.4155/fmc.11.23.

[29] Meng Li et al. “Shapley Value: From Cooperative Game to Explainable Artificial Intelligence”. In: *Autonomous Intelligent Systems* 4.1 (Feb. 9, 2024), p. 2. ISSN: 2730-616X. DOI: 10.1007/s43684-023-00060-8. URL: https://doi.org/10.1007/s43684-023-00060-8 (visited on 12/11/2025).

[30] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. Nov. 25, 2017. DOI: 10.48550/arXiv.1705.07874. arXiv: 1705.07874 [cs]. URL: http://arxiv.org/abs/1705.07874 (visited on 11/04/2025). Pre-published.

[31] *Machine Learning*. In: *Wikipedia*. Feb. 3, 2025. URL: https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1273675406 (visited on 02/06/2025).

[32] Ričards Marcinkevičs and Julia E. Vogt. “Interpretable and Explainable Machine Learning: A Methods-Centric Overview with Concrete Examples”. In: *WIREs Data Mining and Knowledge Discovery* 13.3 (2023), e1493. ISSN: 1942-4795. DOI: 10.1002/widm.1493. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1493 (visited on 02/21/2025).

[33] Riccardo Miotto et al. “Deep Learning for Healthcare: Review, Opportunities and Challenges”. In: *Briefings in Bioinformatics* 19.6 (May 6, 2017), pp. 1236–1246. ISSN: 1467-5463. DOI: 10.1093/bib/bbx044. PMID: 28481991. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6455466/ (visited on 02/06/2025).

[34] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Jan. 27, 2020, pp. 607–617. DOI: 10.1145/3351095.3372850. arXiv: 1905.07697 [cs]. URL: http://arxiv.org/abs/1905.07697 (visited on 09/10/2025).

[35] Meike Nauta et al. “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI”. In: *ACM Comput. Surv.* 55 (13s July 13, 2023), 295:1–295:42. ISSN: 0360-0300. DOI: 10.1145/3583558. URL: https://dl.acm.org/doi/10.1145/3583558 (visited on 02/21/2025).

[36] *Nos Valeurs – Parti Conservateur Du Québec*. URL: https://www.conservateur.quebec/parti/nos-valeurs/ (visited on 10/08/2025).

[37] *Notion – The all-in-one workspace for your notes, tasks, wikis, and databases.* Notion. URL: https://www.notion.so (visited on 02/06/2025).

[38] *Page Non Trouvée – Parti Conservateur Du Québec*. URL: https://www.conservateur.quebec/liberte_22_economie?utm_source=chatgpt.com (visited on 10/08/2025).

[39] *Parti Québécois*. In: *Wikipedia*. Sept. 15, 2025. URL: https://en.wikipedia.org/w/index.php?title=Parti_Qu%C3%A9b%C3%A9cois&oldid=1311396573 (visited on 10/08/2025).

[40] Maximilian Pichler and Florian Hartig. “Machine Learning and Deep Learning—A Review for Ecologists”. In: *Methods in Ecology and Evolution* 14.4 (2023), pp. 994–1016. ISSN: 2041-210X. DOI: 10.1111/2041-210X.14061. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14061 (visited on 12/12/2025).

[41] *Québec Solidaire*. In: *Wikipedia*. Sept. 28, 2025. URL: https://en.wikipedia.org/w/index.php?title=Qu%C3%A9bec_solidaire&oldid=1313934325 (visited on 10/08/2025).

[42] *Résultats des élections générales*. Élections Québec. Nov. 25, 2021. URL: https://www.electionsquebec.qc.ca/resultats-et-statistiques/resultats-generales/2022-10-03/ (visited on 10/13/2025).

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Aug. 9, 2016. DOI: 10.48550/arXiv.1602.04938. arXiv: 1602.04938 [cs]. URL: http://arxiv.org/abs/1602.04938 (visited on 12/11/2025). Pre-published.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 25, 2018). ISSN: 2374-3468. DOI: 10.1609/aaai.v32i1.11491. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11491 (visited on 09/10/2025).

[45] Benedek Rozemberczki et al. *The Shapley Value in Machine Learning*. May 26, 2022. DOI: 10.48550/arXiv.2202.05594. arXiv: 2202.05594 [cs]. URL: http://arxiv.org/abs/2202.05594 (visited on 12/11/2025). Pre-published.

[46] Cynthia Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. Sept. 22, 2019. DOI: 10.48550/arXiv.1811.10154. arXiv: 1811.10154 [stat]. URL: http://arxiv.org/abs/1811.10154 (visited on 12/12/2025). Pre-published.

[47] Cynthia Rudin et al. *Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges*. July 10, 2021. DOI: 10.48550/arXiv.2103.11251. arXiv: 2103.11251 [cs]. URL: http://arxiv.org/abs/2103.11251 (visited on 07/09/2025). Pre-published.

[48] Cynthia Rudin et al. "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges". In: *Statistics Surveys* 16 (none Jan. 2022), pp. 1–85. ISSN: 1935-7516. DOI: 10.1214/21-SS133. URL: https://projecteuclid.org/journals/statistics-surveys/volume-16/issue-none/Interpretable-machine-learning-Fundamental-principles-and-10-grand-challenges/10.1214/21-SS133.full (visited on 02/07/2025).

[49] Cynthia Rudin et al. *Amazing Things Come From Having Many Good Models*. July 10, 2024. DOI: 10.48550/arXiv.2407.04846. arXiv: 2407.04846 [cs]. URL: http://arxiv.org/abs/2407.04846 (visited on 12/11/2025). Pre-published.

[50] Lesia Semenova, Cynthia Rudin, and Ronald Parr. "On the Existence of Simpler Machine Learning Models". In: *2022 ACM Conference on Fairness Accountability and Transparency*. June 21, 2022, pp. 1827–1858. DOI: 10.1145/3531146.3533232. arXiv: 1908.01755 [cs]. URL: http://arxiv.org/abs/1908.01755 (visited on 12/13/2025).

[51] *Sparsity - an Overview | ScienceDirect Topics*. URL: https://www.sciencedirect.com/topics/computer-science/sparsity (visited on 02/07/2025).

[52] *What Is Deep Learning? | IBM*. June 17, 2024. URL: https://www.ibm.com/think/topics/deep-learning (visited on 02/06/2025).

[53] *What Is Machine Learning (ML)? | IBM*. Sept. 22, 2021. URL: https://www.ibm.com/think/topics/machine-learning (visited on 02/06/2025).

[54] Ian R. White, Patrick Royston, and Angela M. Wood. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice". In: *Statistics in Medicine* 30.4 (2011), pp. 377–399. ISSN: 1097-0258. DOI: 10.1002/sim.4067. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4067 (visited on 03/15/2025).

[55] Mengwei Xu et al. "A First Look at Deep Learning Apps on Smartphones". In: *The World Wide Web Conference*. WWW '19: The Web Conference. San Francisco CA USA: ACM, May 13, 2019, pp. 2125–2136. ISBN: 978-1-4503-6674-8. DOI: 10.1145/3308558.3313591. URL: https://dl.acm.org/doi/10.1145/3308558.3313591 (visited on 02/07/2025).

[56] Yujia Zhang et al. *"Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations*. June 4, 2019. DOI: 10.48550/arXiv.1904.12991. arXiv: 1904.12991 [cs]. URL: http://arxiv.org/abs/1904.12991 (visited on 12/11/2025). Pre-published.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc