Machine Learning (2025) 114:225
https://doi.org/10.1007/s10994-025-06847-5

Region-aware Minimal Counterfactual Rules for Model-agnostic Explainable Classification

Guido Gagliardi[1,3,4] · Antonio Luca Alfeo[7] · Riccardo Guidotti[5,6] · Mario G. C. A. Cimino[1,2]

Received: 24 February 2025 / Revised: 17 June 2025 / Accepted: 18 July 2025 /
Published online: 4 September 2025

Abstract
The increasing demand for transparency in machine learning has spurred the development of techniques that provide faithful explanations for complex black-box models. In this work, we introduce RaMiCo (Region Aware Minimal Counterfactual Rules), a model-agnostic method that extracts global counterfactual rules by mining instances from diverse regions of the input space. RaMiCo focuses on single-feature substitutions to generate minimal and region-aware rules that encapsulate the overall decision-making process of the target model. These global rules can be further localised to specific input instances, enabling users to obtain tailored explanations for individual predictions. Comprehensive experiments on multiple benchmark datasets demonstrate that RaMiCo achieves competitive fidelity in replicating black-box behaviour and exhibits high coverage in capturing the intrinsic structure of white-box classifiers. RaMiCo supports the development of trustworthy and secure machine learning systems by providing transparent, human-understandable explanations in the form of concise global rules. This design enables users to verify and inspect the model's decision logic, reducing the risk of hidden biases, unintended behaviours, or adversarial exploitation. These features make RaMiCo particularly suitable for applications where the reliability, safety, and verifiability of automated decisions are essential.

Keywords Explainable artificial intelligence · Rule-based explanations · Counterfactual explanation · Region-aware rule extraction · Model-agnostic explanations

1 Introduction

As machine learning and artificial intelligence (AI) methods become more widespread and complex, it is more and more important to provide user-friendly explanations for their decision-making processes. Explainable AI (XAI) methods generate explanations for the reasoning behind AI predictions, enabling domain experts to validate and trust the AI models and allowing those affected by AI-based decisions to understand their

Editors: Annalisa Appice, Giuseppina Andresini, Przemyslaw Biecek, Christian Wressnegger.

Extended author information available on the last page of the article

rationale (Schoenborn & Althoff, 2019). This need for explainability is reinforced by recent regulations on personal data processing, such as the AI Act and the General Data Protection Regulation (GDPR), which require a certain level of transparency for AI to be used in real-world decision- making (Foulsham et al., 2019).

Also, the integration of AI into real-world critical systems raises significant security concerns about the safety, reliability, and verifiability of automated decisions. In such contexts, it is essential not only to ensure high predictive performance but also to provide mechanisms that make model behaviours transparent, understandable, and resistant to manipulation or adversarial influence. Explanations must therefore serve a dual purpose: they should enable human stakeholders to comprehend and validate the decision process, and they should contribute to the robustness of the system by exposing potential vulnerabilities or unintended behaviours that could compromise its security or trustworthiness.

XAI approaches can be categorised as model-specific methods, tailored to specific machine learning (ML) architectures like neural networks (Alfeo et al., 2023a), or model-agnostic methods (Gagliardi et al., 2023), applicable to any black-box AI model. Explanations can also be global or local (Arrieta et al., 2020): global methods explain the entire decision logic of a model, while local methods focus on decisions for specific input instances. Finally, the explanations can be characterized by different forms: *(i)* instance-based explanations, aimed at linking a given instance to prototypes or counterfactual examples, enabling similarity-based reasoning, e.g. a counterfactual is an instance similar to the one to explain corresponding to a different model outcome (Guidotti, 2024); *(ii)* attribution-based explanations, aimed at assessing each input feature's contribution to the prediction, both on local and global level (Alfeo et al., 2023b); and *(iii)* rule-based explanations, aimed at approximating the model's decision process using logical rules, such as *if-then* statements (Guidotti et al., 2024).

Rule-based and counterfactual explanations are highly user-friendly, as they offer predictive insights into a model's behavior, presenting it in terms of causes and effects (Chou et al., 2022), which mimics human reasoning (Byrne, 2016). Still, counterfactual explanations are inherently local, and rule-based explanations can become verbose and hard to interpret depending on their complexity and numerosity. To address these limitations, researchers are exploring hybrid approaches that combine the strengths of both. For example, Guidotti et al. (2024) introduces a more abstract form of counterfactuals using logical rules instead of simple feature flips, along with an abstract representation of a sample's neighborhood with different black-box outcomes.

These methods simplify complex model behaviors into understandable rules, enabling global and local insights into changes in black-box model outcomes. Many rule-based XAI approaches use surrogate models, such as distillation-based approaches exploiting decision trees, to extract rules able to approximate the behavior of an AI model (Love et al., 2023). While effective in capturing the average decision-making process, proxy models often fail to represent subtle feature variations that result into class changes, which are crucial for counterfactual explanations. Such variations typically occur in specific subgroups of training data, like near decision boundaries or between class prototypes, and are often missed by surrogate models due to their reliance on generalized approximations rather than regional feature dynamics. Existing methods in rule extraction focus on global explainability, often compromising regional explainability of minor subgroups of data to achieve better overall performance across the entire data distribution (Chen et al., 2024).

This paper introduces a novel approach to generate counterfactual rules able to summarise the decision-making process of any black-box model for tabular data. The explanations produced are predictive of the model's behaviour both *locally*, i.e., a rule for each single instances, and globally, i.e., unveiling the entire decision logic focusing on specific regions of the sample space.

The proposed approach, named RaMiCo (Region Aware Minimal Counterfactual Rules), systematically identifies instances from a black-box input space to derive minimal yet highly informative counterfactual rules. By concentrating on single-feature substitutions within designated regions of the input space, RaMiCo produces explanations that are both faithful to the original model's decision-making process and inherently minimal, thereby enhancing interpretability. RaMiCo not only ensures that the derived rules closely reflect the decision-making mechanisms of the underlying classifier but also ensures that the generated counterfactuals are maintained within the natural data distribution.

Consequently, RaMiCo specifically addresses the need for transparent, verifiable, and manipulation-resistant machine learning models in critical decision-making scenarios by providing an explanation framework that extracts concise, semantically meaningful global rules from black-box models. Its design supports the creation of transparent, human-readable explanations that allow end-users and domain experts to understand and verify the decision logic of ML systems, thereby reducing the risks of hidden biases or unintended behaviours. Moreover, the ability of RaMiCo to produce consistent and inspectable rules facilitates system auditing and resilience analysis, contributing to the development of secure and reliable AI solutions.

RaMiCo is firstly evaluated using a white-box model, such as a decision tree, thereby illustrating its capacity to comprehensively extract global minimal rules that reflect the intrinsic structure of the tree, i.e., the rules *coverage* of the tree. RaMiCo is also assessed in the extraction of global rules from black-box models, demonstrating the coherence of these rules with the problem input space, i.e., the rules *plausibility*, and their strong alignment with the model's decision-making processes, i.e., the rules *fidelity*.

The rest of the paper is organized as follows. Section 2 outlines the proposed XAI approach, Sect. 3 describes the experimental dataset, Sect. 4 details the experimental setup and results, and Sect. 5 discusses the findings and presents the conclusions.

2 Related works

Rule-based XAI methods (Rapp et al., 2024) aim to create simpler and interpretable rule-based models that are able to approximate the behavior of an AI black-box model (Mahya & Fürnkranz, 2023). These methods can be categorized as follows (Marshakov, 2021):

- Pedagogical: These methods treat the ML model as a black box, focusing on its learned function. They extract rules that describe the relationship between inputs and outputs without using internal model details (De Fortuny & Martens, 2015).
- Decompositional: These methods analyze the model's internal structure, such as weights or activation patterns in neural networks, to break down decision-making

into transparent components, explaining how specific features contribute to predictions (Zarlenga et al., 2021).
- Eclectic: These hybrid methods combine black-box analysis with internal model inspection to extract rules describing input–output relationships (Hao et al., 2022).

In this context, model-agnostic post-hoc XAI methods (Mahya & Fürnkranz, 2023) belong to the pedagogical category. Many use decision tree-based proxy models, treating the network as a black box and training proxies on input data and the AI model's predictions. For instance, the SIRUS approach (Bénard et al., 2021) is a rule-based XAI approach working into two steps: training a black-box model and constructing a secondary model to extract non-overlapping rules that capture robust patterns in the data. As another example, the inTrees approach (Deng, 2019) can extract, prune, and summarize rules from tree ensembles like Random Forests and Boosted Trees, ranking rules based on their length, support, and error. As such this method does not fall exactly under the umbrella of the model-agnostic approaches. Finally, well-known methods like LIME (Salih et al., 2024) and Anchor (Elkhawaga et al., 2023) provide only local explanations and require continuous features to be discretized before use. The approach proposed in Luo et al. (2020) shares similar limitations. It uses association rule mining to transform rule extraction into identifying frequent item sets–groups of items that commonly appear together in training samples. This requires numerical features to be discretized into categorical items (Chakraborty et al., 2020), leading to a trade-off between speed and performance (Ley et al., 2022). Moreover, discretizing features individually, based solely on their marginal distributions rather than joint distributions, assumes feature independence–an assumption often invalid in real-world scenarios.

Additionally, evaluating rule sets involves balancing support and confidence. While proxy methods effectively capture the average decision-making process of the model, they struggle to account for subtle feature variations that lead to class changes (Chen et al., 2024). These variations are often exhibited in specific subgroups of training samples, such as those near decision boundaries or between different classes' prototypes. As such, these differences are poorly represented by surrogate-based rule extraction due to its reliance on generalized approximations rather than precise local feature dynamics (Freiesleben & König, 2023).

To address this issue, the authors in Chen et al. (2024) introduce a model-agnostic method for extracting rules from specific data subgroups, including automatic rule generation for numerical features, improving regional explainability in machine learning models. This approach aligns with XAI research focusing on XAI for counterfactual rules. As an example, Rawal and Lakkaraju (2020) introduced Actionable Recourse Summaries (AReS), a framework that uses a greedy heuristic search to generate global counterfactual rules, though it may lead to unstable and inaccurate results. Similarly, the Regional Counterfactual Rule (Amoukou & Brunel, 2022) also targets global counterfactual rules but directly handles continuous features and leverages random forest partitions

to reduce the search space, focusing on high-density regions to ensure plausibility. More recently, authors in Guidotti et al. (2024) proposed a method to generate factual and counterfactual rules using a genetic algorithm to synthesize data similar to the instance being explained. The approach builds an ensemble of local decision trees via bagging, which are then merged into a single interpretable decision tree for stability. Factual rules are derived from the root-to-leaf path of the decision tree, while counterfactual rules are obtained by identifying minimal feature changes that alter the prediction. However, merging decision trees and generating genetic neighborhoods can be computationally intensive and sensitive to hyperparameter settings. Despite these challenges, the method is modern and effective (Guidotti et al., 2024) and will be considered as a competitor in this study.

Our proposal provides an effective, post-hoc and model-agnostic XAI approach for rule extraction. Compared to similar solutions introduced in this section, the proposed approach avoids any transformation of categorical features (Chakraborty et al., 2020) and complex rule search procedures (Rawal & Lakkaraju, 2020; Guidotti et al., 2024). Above all, our approach overcomes the trade-off between global and local representation of the decision-making process of the ML model (Freiesleben & König, 2023; Chen et al., 2024). Indeed, by mapping different sub-populations of the training instances, the proposed approach is able to derive global counterfactual rules while also offering a good approximation of local decision-making of the ML model.

3 Methodology

In this section we present the RaMiCo (Region Aware Minimal Counterfactual Rules) XAI method, a model-agnostic framework aimed at extracting minimal counterfactual rules from a black-box classifier. With *minimal*, we refer to rules characterized by the least number of items i.e., features to change in order to flip the output of the model.

Given a classification problem where $C = \{0, 1, \dots\}, c_i \in \mathbb{N}$ is the class set, and $M : X \to C$ is a black-box model, with X *samples space* of the classification problem, $X \subseteq R^n$, with $n \in \mathbb{N}$ features that characterize the samples within X. We say that the model M assigns (or predicts) a class, $c \in C$, to each sample $x \in X$ as $M(x) = c$.

RaMiCo analyzes both the class input space $X_c \subset X$, i.e. the partition of samples associated with class c, as well as the input spaces outside the class, i.e. $X_{\neq c} : X_c \cup X_{\neq c} \equiv X$ and $X_c \cap X_{\neq c} \equiv \emptyset$. Drawing from X_c, RaMiCo selects s samples, referred to as the *factuals*. For each of these s samples, it extracts k samples from $X_{\neq c}$, designated as *counterfactuals*. By capitalizing on the distinctions between each factual instance and its counterfactuals, RaMiCo derives global rules that elucidate the model's decision-making processes.

For each class space X_c present in X, the pseudocode of the proposed framework is detailed in Algorithm 1 and works according to the following steps:

Algorithm 1 RaMiCo: Region Aware Minimal Counterfactual Rules

Require: Black-box model M, sample space $X \subseteq \mathbb{R}^n$, set of classes $\mathcal{C}$, number of factual samples s, number of counterfactuals k

for each class $c \in \mathcal{C}$ **do**
  Let $X_c \subset X$ be the set of samples belonging to class c
  Let $X_{\neq c} = X \setminus X_c$ ▷ All samples not in class c
  Select s samples from $X_c : X_c^{(s)}$
  for each selected sample $x_i \in X_c^{(s)}$ **do**
    $X_{\neq c}^{(s)_i} \leftarrow$ mine_counterfactuals(x_i, k, $X_{\neq c}$)
  end for
  for each pair (x_i, x_j) with $x_i \in X_c^{(s)}$ and $x_j \in X_{\neq c}^{(s)_i}$ **do**
    Identify all the attributes (features) $a : x_i[a] \neq x_j[a]$
    $G_{x_i,x_j} \leftarrow \emptyset$
    for each identified attribute a **do**
      Generate $x^* = x_i$
      $x^*[a] \leftarrow x_j[a]$
      $G_{x_i,x_j} \leftarrow x^*$
    end for
  end for
  Let $\mathbf{G} = \bigcup_{(x_i,x_j)} G_{x_i,x_j}$ ▷ Union of all generated sets
  Select $\mathbf{G}_{q'} \subset \mathbf{G}$ with $x \in \mathbf{G}_{q'} : M(x) \neq c$
  for each sample $x^* \in \mathbf{G}_{q'}$ **do**
    Let $p = M(x^*), p \neq c$ predicted class of x^*
    Create a *rule-record* of the form:
      (c, p) by switching attribute a from $x_i[a]$ to $x_j[a]$
  end for
  Aggregate the rule-records (RR) that share the same starting class c, switching class p, and attribute a
  $RR_{c,p,a} = \bigcup_{c,p,a}(c, p), a : x_i[a] \rightarrow x_j[a]$
  for each set $RR_{c,p,a}$ **do**
    Generate the final rule considering the expected values and standard deviation of $x_i[a]$ and $x_j[a]$ over the $RR_{c,p,a}$ set:
      $(c, p), a : \mathbb{E}[x_i[a]] \pm \sigma_i[x_i[a]] \rightarrow \mathbb{E}[x_j[a]] \pm \sigma_j[x_j[a]]$
  end for
end for

1. Select s samples from X_c, $X_c^{(s)}$, and, for each of these samples, i, mine its k counterfactuals from $X_{\neq c}$, $X_{\neq c}^{(s)_i}$.
2. For each of the $k \times s$ couples of samples, (x_i, x_j), generate a set of new samples G_{x_i,x_j} by switching a single feature value $a : x_i[a] \neq x_j[a]$ between x_i and x_j.
3. Select $\mathbf{G}_{q'}$ from $\mathbf{G}$ by evaluating the output of the model M on it.
4. For each sample in $\mathbf{G}_{q'}$, consider its predicted class p, and generate a rule-record as: (c, p) by switching attribute a from $x_i[a]$ to $x_j[a]$
5. Generate a rule from the rule-records by aggregating them by the same starting class c, switching class p and target feature a and computing the average and

the standard deviation of the values $x_i[a]$ and $x_j[a]$ during the aggregation: $c \to p$, $a : \mathbb{E}[x_i[a]] \pm \sigma_i[x_i[a]] \to \mathbb{E}[x_j[a]] \pm \sigma_j[x_j[a]]$

Suppose, for instance, that three factual–counterfactual rule-records are mined for the transition from a class 0 to class 1 on feature `age`:

$$(0 \to 1,\ \text{age}) :\ 30 \to 35, \quad (0 \to 1,\ \text{age}) :\ 32 \to 38, \quad (0 \to 1,\ \text{age}) :\ 28 \to 33.$$

RaMiCo fixes the key $(0 \to 1,\ \text{age})$ and computes

$$\bar{v}_1 = \frac{30 + 32 + 28}{3} = 30, \quad \sigma_1 = \sqrt{\frac{(30-30)^2 + (32-30)^2 + (28-30)^2}{3}} \approx 2,$$

$$\bar{v}_2 = \frac{35 + 38 + 33}{3} = 35.3, \quad \sigma_2 = \sqrt{\frac{(35-35.3)^2 + (38-35.3)^2 + (33-35.3)^2}{3}} \approx 2.05.$$

The resulting global rule is then presented as:

$$\text{If age increases from } 30 \pm 2 \text{ to } 35.3 \pm 2.05, \quad \text{then class } 0 \to 1.$$

It is important to note that through this aggregation, RaMiCo effectively prevents conflicts within rules, as there exists a singular rule associated with the triplet c, p, a, consisting of the starting class, ending class, and feature. If this triplet refers to rule-records that exhibit excessive heterogeneity or conflict, this will manifest as an increase in the standard deviation corresponding to the rule. In situations where the standard deviation surpasses the acceptable threshold for the user, it is advisable to separate the rule records and perform the rule aggregation again until an acceptable standard deviation is achieved.

The cardinality of G_{x_i,x_j} is equal to the number of mismatches in the value of the features between x_i and x_j. Consider the overall group of sets $\mathbf{G}$ that includes all the sets generated for each couple of samples.

To select these instances different regions of the input space are exploited. RaMiCo employs the *mine_counterfactuals* function (Algorithm 1, line 6). This function represents a sequence of combined mining strategies. These strategies include:

- **Decision-Boundary Mining**: Factuals and counterfactuals are identified in close proximity between X_c and $X_{\neq c}$. The extracted rules highlight minimal feature differences that change the classification outcomes. This mining strategy exploits the pairwise distance between X_c and $X_{\neq c}$ and selects the s samples from X_c with minimal distance to $X_{\neq c}$. Once the *factuals* have been mined, the *counterfactuals* are selected as the first k samples with the minimal distance to each *factual*.
- **Prototype Mining**: Factual and counterfactuals are selected as centroids for the class and out-class spaces. The extracted rules denote significant feature differences altering classification outcomes. This mining strategy computes the s-medoids of the X_c space as the *factuals* and the k-medoids of the $X_{\neq c}$ space as *counterfactuals*.
- **Random Mining**: Factual and counterfactuals are randomly chosen from within and outside the class space. The extracted rules comprehensively cover the sample space, offering insights from various samples. However, the learned rules are less interpretable as they do not originate from a specific space section.

Since we deal with mixed categorical and numerical features values, the distance is computed as the Gower distance, defined as:

$$d(x_i, x_j) = \frac{1}{M}\sum_{m=1}^{M} s_{ij}^{(m)} \begin{cases} (1) s_{ij}^{(m)} = \frac{|x_i^{(m)} - x_j^{(m)}|}{max(x^{(m)}) - min(x^{(m)})} \\ (2) \begin{cases} s_{ij}^{(m)} = 1, \leftarrow x_i^{(m)} = x_j^{(m)} \\ s_{ij}^{(m)} = 0, \leftarrow x_i^{(m)} \neq x_j^{(m)} \end{cases} \end{cases}$$

where equation (1) is for numerical features and (2) is for categorical features.

3.1 Features values switching

The generation of the set G_{x_i,x_j} can produce outliers due to switching feature values in different space regions. The inferred rules may transition between feature values that are unfeasible for space X.

For instance, if the sample x_j possesses attribute a such that $x_j[a] = v_2 \neq x_i[a]$, where $x_i[a] = v_1$, the algorithm generates a new sample x^* defined as $x^* = x_i$ with $x^*[a] = v_2$. If the model's prediction for x^*, denoted as $M(x^*)$, results in $p \neq c$, the inferred rule is represented as $(c, p), a : v_1 \rightarrow v_2$.

However, this inference may be problematic if $x^* \notin X$ or if it represents an implausible feature combination, such as one that contradicts existing feature dependencies. In such cases, the inferred rule suggests a transition to v_2, a value that may not be valid within the actual data distribution. Moreover, even if x^* does not violate feature dependencies, it might correspond to a rare or outlier instance in the dataset. As a result, while the inferred rule remains faithful to the model–since $M(x^*) = p \neq c$–it may not accurately capture a general trend within the input data, depending on the likelihood of x^* occurring within X. To address this issue, RaMiCo employs an isolation forest model on X to assess sample coherence from G_{x_i,x_j} for X. The sample's plausibility, ρ, i.e., the isolation forest's outlier score, is assigned to each rule record and averaged in the aggregation step.

$$\mathbb{E}_{RR_{c,p,a}}[\rho] = \rho_a$$

This way, each rule has a plausibility score assigned to it and can be read as:
$(c, p), a : \mathbb{E}[x_i[a]] \pm \sigma_i[x_i[a]] \rightarrow \mathbb{E}[x_j[a]] \pm \sigma_j[x_j[a]]$ with plausibility ρ_a
Or more compacted:
$(c, p), a, \rho_a : \mathbb{E}[x_i[a]] \pm \sigma_i[x_i[a]] \rightarrow \mathbb{E}[x_j[a]] \pm \sigma_j[x_j[a]]$

3.2 Computational complexity

Let n denote the number of input features, s the number of factual samples, and k the number of counterfactual samples. If the problem has m instances, we assume that s and k are small enough compared to m. Hence, the mining step selects a small number of instances highly representative of the input space, so that $s \times k \leq m$. Note that, if this assumption doesn't hold, in the worst case we would have $s = k = m$, hence $s \times k = m^2$

We also assume that each prediction of the model (that is, the evaluation in one instance) requires constant time $O(1)$, as its computational cost depends on the internal structure of the model.

In total, we have $s \times k$ factual-counterfactual pairs, and for each pair, we need to evaluate if, by switching their feature values, the model changes its prediction. First, we check if the feature values $s_i, k_i, i \in 1 \dots n$ are different $s_i \neq k_i$, and in case they are, we evaluate the model on each new generated instance. In the worst case, all the n features are different between the factual and the counterfactual pair, so we need $O(n)$ operations to check the difference, and $O(n) \times O(1) = O(n)$ model predictions. The total cost of this step is $O(n) + O(n) = 2 \cdot O(n) = O(n)$ for each factual counterfactual pair $k \times s \times O(n)$. If our starting assumption $s \times k \leq m$ holds the computational cost is $O(m \cdot n)$ or $O(m^2 \cdot n)$ in case it doesn't.

In RaMiCo we utilise 3 mining approaches to select the factual and counterfactuals from the input space: Random, Prototype and Decision Boundary. The complexity of the Random approach is $O(k \cdot s)$; if we choose to implement the Prototype method with a $k - medoids$ approach its complexity is $O(k \cdot s \cdot m^2)$; and the complexity of the Decision Boundary approach is $O(m^2)$ due to the distance matrix computation.

Overall, if we keep k, s much smaller compared to m, such as $k \times s \ll m$ we can conclude that the RaMiCo complexity depends mostly on the pairs selection and its complexity is $O(n) + O(m^2)$, if we consider $n \leq m$ than the complexity is $O(m^2)$. If our starting assumption holds, $k \times s \leq m$, then the complexity raise to $O(m \cdot n) + O(m^2) = O(m^2)$. If our starting assumption doesn't hold, we don't have the mining overhead as we select all the samples as factuals and counterfactuals, and the overall complexity is $O(m^2 \cdot n) = O(m^3)$.

Figure 1 illustrates the runtime of RaMiCo's global rule learning process as the number of input samples m increases, with fixed parameters $k = 10$ and $s = 10$ and syntetically generated datasets with $n = 100$ features. The datasets generation process included creating feature clusters that represent distinct classes, $c = 2$, by placing points around the vertices of a high-dimensional space; then introducing dependencies by deriving some features as linear combinations of the informative ones; and finally adding random noise features to increase complexity. This procedure has been implemented using the *make_classification* functionality of the *sklearn* Python library with default parameters. All the simulations run

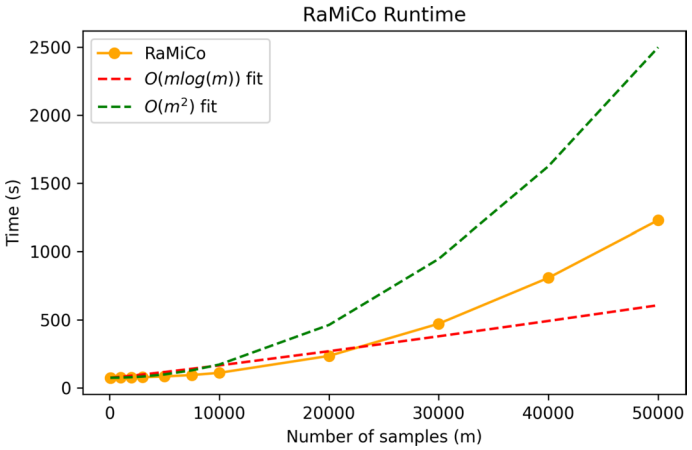

Fig. 1 Runtime of RaMiCo's global rule learning procedure as a function of the number of samples m, with parameters $k = 10$ and $s = 10$. The starting hypothesis $k \times s \leq m$ holds in this simulation. The empirical runtime is compared to theoretical complexity fits of $O(m \log m)$ and $O(m^2)$. The results show that the theoretical upper bound $O(m^2)$ overestimates the actual complexity, which approaches this bound asymptotically for large m

on a 2x24-core AMD Epyc 7402 CPU 2.8 GHz with 256 GB of RAM. Note that in this case, the starting hypothesis $k \times s \leq m$ holds, hence the expected theoretical complexity is $O(m^2)$. The empirical results are compared to theoretical complexity curves of $O(m \log m)$ and $O(m^2)$. As expected, the actual runtime of RaMiCo grows faster than $O(m \log m)$ but remains below the quadratic bound $O(m^2)$ in the range of sample sizes tested. This indicates that while the theoretical complexity of $O(m^2)$ provides a conservative upper bound, the practical runtime behaviour of RaMiCo is more efficient in typical scenarios, approaching the upper bound only asymptotically for large m.

Note that this computational cost needs to be accounted for only once, as the rules are learned globally. The localisation step requires instead to locate the query instance in the rule-records space. If all the factuals generated at least one rule, this space has $s \leq m$ instances. The localisation step then requires $s \cdot log(s)$ operations, or $O(m \cdot log(m))$ operations if our starting assumption doesn't hold.

If we consider evaluating RaMiCo on a set of m new instances, we can split the global rule learning cost to each new instance during the localisation. In this case each instance localisation cost would be $\frac{O(m^2) * O(s \cdot log(s))}{O(m)} = O(m) \cdot O(s \cdot log(s))$. If we consider $k \times s \leq m, s \ll m$ the computational cost of RaMiCo in evaluation is linear $O(m)$, while if our starting assumption doesn't hold, i.e. $s \simeq m$, then its complexity becomes $O(m^2 \cdot log(m))$.

3.3 From global to local

RaMiCo derives minimal counterfactual rules from a black-box classifier. These rules elucidate the decision-making process at a global model level. However, applying such global rules to specific input instances can be challenging, as the rules are valid for the regions of the sample space from which they have been extracted, i.e., we say that the rules are *active* in such regions.

It's worth noticing that RaMiCo relies on single-feature substitutions during rule-learning; thanks to this, RaMiCo is capable to learn a set of minimal rules maximising rule comprehensibility. On the other hand, given the fact that it doesn't rely on multi-feature transitions, we are not sure of tracking the model class transition between the factual and counterfactual pair. On the other hand, relying on multi-feature transitions implies having a greater number of terms in the rule chain, hence decreasing the rule comprehensibility.

The localisation step allows RaMiCo to implicitly capture multi-feature interactions, as the set of activated rules reflects the joint behaviour of features within that local context. This mechanism ensures that local explanations remain expressive and context-aware, preventing excessive simplification even though the global rule base is composed of individually simple components.

To identify the minimal rules relevant for a query sample $x \in X$, RaMiCo maps the queries within its learned rule space and provides a set of rules *active* for it. To this aim, RaMiCo employs a NearestNeighbours (NN) model trained on the selected *factuals*. The NN model replies to any query sample $x \in X$ with a *factual* and its generated rule-records from nearby factuals to x. This way, RaMiCo is able to exploit the global rule records and provide local rules for a given input sample x.

Formally, let $\mathbf{RR}$ be the union of all rule records $RR_{c,p,a}$ extracted from RaMiCo at the end of Algorithm 1. The system assigns each rule record rr, defined as $rr : (c, p), a : x_i[a] \rightarrow x_j[a]$ with $rr \in \mathbf{RR}$, its corresponding factual sample x_i. It then constructs the Localization Set $\mathbf{LS}$, which consists of pairs $ls = (rr, x_i)$ such that $ls \in \mathbf{LS}$.

Let r be the user-defined number of rule records to consider, and let $x \in X$ be a query sample. The localization system returns the r-nearest records in **LS** to x, where the distance between elements of X and elements of **LS** is based on the factual elements of the localization space:

$$x \in X, \quad ls \in \mathbf{LS}, \quad ls = (rr, x_i)$$

$$\text{dist}(x, ls) = \text{dist}(x, x_i)$$

At inference time, when the user queries the global active rules for a sample x, the system retrieves the r-nearest records from **LS** using the defined distance metric. It then aggregates these records by class transition and feature (c, p, a), computing the mean transition value:

$$\mathbb{E}_{c,p,a}[x_j[a]] = \mathbf{x_j}[\mathbf{a}]$$

Thus, the system provides the active rule:

$$(c, p), a : x[a] \rightarrow \mathbf{x_j}[\mathbf{a}]$$

If multiple features or class transitions are present among the r-nearest rule records, the system generates multiple active rules.

4 Experiments

In this section, we provide a comprehensive evaluation of our proposed method, assessing its performance across various datasets and machine learning models. We compare our approach against state-of-the-art explainable AI (XAI) methods for rule extraction and analyze the results using well-defined evaluation metrics.

The section is structured as follows:

- Experimental Setting: We describe the experimental setup and introduce the different experiments whose results are discussed in the following sections.
- Datasets and Black-Box Models: We present the datasets used for evaluation and specify the black-box models employed in the experiments.
- Competitors: We outline the explainability methods used as baselines for comparison.
- Evaluation Metrics: We define the quantitative metrics used to assess the effectiveness and reliability of the explanations.
- Results: We analyze and discuss the experimental performance of our approach based on the established setup.

4.1 Experimental setting

To evaluate the effectiveness of our proposed method, we designed two distinct experiments aimed at assessing the quality of the extracted rules in different scenarios.

In the first experiment, we trained a white-box classifier, specifically a Decision Tree (DT), and applied RaMiCo to extract decision rules from it. Since the internal structure of a decision tree is inherently interpretable, this setting allows us to directly compare the extracted rules with the actual decision paths within the tree. The objective is to measure

Table 1 Summary of the datasets used in this study, detailing the number of features, categorical features, and total samples for each dataset

Dataset	Features	Categorical Features	Number of Samples	DT (%)	MLP (%)	XGBoost (%)
Telco	19	16	7032	78	74	77
Adult	14	8	30,162	85	82	87
Car Evaluation	6	6	1728	93	99	98
Res	17	7	382	90	95	94
Bulk	17	7	440	84	90	88%

how well RaMiCo captures the model's true decision logic, providing a baseline for evaluating its rule extraction capabilities.

In the second experiment, we evaluated rule extraction methods in a more challenging setting by training two different black-box classifiers: XGBoost, a gradient-boosting decision tree model known for its strong predictive performance, and Multilayer Perceptron (MLP), a neural network model capable of capturing complex, nonlinear relationships. To interpret these models, we applied three different explainable AI approaches for rule extraction, including RaMiCo, to extract decision rules. We then compared the extracted rules to assess their effectiveness in explaining the decision-making process of each black-box model.

4.1.1 Datasets and black box models

This section details the datasets used to evaluate RaMiCo against leading rule extraction methods, including 2 benchmark datasets from the UCI repository: *Adult*[1] and *Car Evaluation*;[2] IBM's *Telco customer churn*[3] dataset; and a real-world dataset from a real-world tissue manufacturing facility.

This facility manufactures industrial equipment used for the production of tissue paper. Each machine is composed of two main parts: the embosser and the rewinder. The rewinder unwinds and layers reels of raw paper, subsequently transferring them to the embosser. The embosser employs rubber and steel rollers to compress and adhere the tissue layers while embossing a pattern onto the paper. Each unit is tested with various types of paper and diverse production parameters, including rewinder speed and embossing pressure. Measurements are taken from the final product for each configuration. As with many other datasets, the company's data has numerous missing values and is subjected to specific preprocessing steps. Columns and rows with more than 50% missing data are removed. The remaining data instances are grouped based on complete categorical features. Numerical missing values are replaced with the median of the respective cluster, whereas categorical ones use the mode. A detailed description of these datasets can be found at (Alfeo et al., 2023a).

[1] https://archive.ics.uci.edu/dataset/2/adult.
[2] https://archive.ics.uci.edu/dataset/19/car+evaluation.
[3] https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=samples-telco-customer-churn.

Table 1 summarizes the datasets, including feature distributions, sample sizes, and the predictive performance, i.e. accuracy on the test set, of the three classifiers used in our experiments.

4.1.2 Competitors

Two competitors have been identified to compare RaMiCo with: a Trepan-inspired approach, which extracts minimal rules by training a surrogate decision tree of the black box-model, and Local Rule-Based Explanation (LORE) (Guidotti et al., 2024).

LORE creates local counterfactual rules by analyzing individual samples and identifying changes needed to alter predictions. It computes a counterfactual for each query and derives rules from the feature differences between the query and its counterfactual.

The Trepan-inspired approach, which we refer as *Minimal Rule Tree* (MRT), consists of training a surrogate decision-tree classifier on the black box model under inspection. Then the approach extract minimal rules from the trained decision tree extracting its leaf nodes.

MRT extracts minimal decision rules from DTs by analyzing how slight changes in input features affect the model's prediction. It starts with an empty set of rules and retrieves the tree's root node. For each sample in the dataset, the algorithm first determines its predicted class by the decision tree, then identifies the specific leaf node that the sample reaches. From this leaf node, it extracts the relevant feature and its threshold value. The sample is then slightly modified by increasing the value of that feature by its standard deviation, as in Pornprasit et al. (2021), Awal and Roy (2024), creating a new sample. If the prediction for the modified sample differs from the original prediction, a minimal rule is recorded that captures the transition from the original class to the new class when the feature value exceeds the threshold.

4.1.3 Evaluation metrics

As outlined by recent surveys (van der Waa et al., 2021; Love et al., 2023), to address the main limitations of rule-based XAI approaches, the following desirable properties are crucial: (i) compatibility with all ML methods, avoiding any assumptions and precondition about the ML model's construction, training, or data; (ii) ability to produce accurate rules describing the ML model behavior (e.g., prediction change); (iii) generation of compact, readable rules by eliminating irrelevant ones while preserving essential information; and (iv) reduced computational cost for rule extraction.

In this regard, a rule-based XAI approach can be evaluated according to the following properties (Awal & Roy, 2024):

- Coverage: how well the extracted rules resemble the inner structure of a white-box classifier, such as a decision-tree, i.e. how much the rules cover the tree. Given a decision tree DT, the coverage is computed as the percentage of leaf-nodes in the tree, for which there exists a rule with the same feature transition $x_i[a] \rightarrow x_j[a]$.

$$\text{Coverage} = \frac{\left|\{r \in R \mid \exists n \in \mathcal{L}(DT) \text{ s.t. } r_a = n_a\}\right|}{|\mathcal{L}(DT)|} \times 100$$

  where: R is the set of extracted rules; $\mathcal{L}(DT)$ represents the set of leaf nodes in the decision tree DT; r_a denotes the feature transition in the extracted rule r; n_a represents the feature transition associated with the leaf node n in the decision tree.

- (In)Comprehensibility, or Minimality: the ability to understand the extracted rules, as rules become more complicated in terms of numbers of logical operations in the rule chain. The (In)Comprehensibility $IC(\cdot)$ of a rule is computed as the number of logical operation (*and*, *or*) in the rule. Note that this is equivalent to the number of features or attributes a involved in the rule chain.
- Fidelity: how well the extracted rules are in mimicking the black-box model behavior. It is computed as, given an input space X, a set of rules S, a matching function $m : X \to S$ associating each sample $x \in X$ to a rule $R : c \to p,\ x_i[a] \to x_j[a]$ active for it, and a black box model M: the percentage number of the samples in X for which $M(x) = c$ and $M(x') = p$ with $x' \equiv x$ apart from $x'[a] \equiv x_j[a]$

$$\text{Fidelity} = \frac{|\{x \in X \mid M(x) = c \text{ and } M(x') = p\}|}{|X|} \times 100$$

The fidelity metric evaluates the alignment of the model's responses to a comprehensive and collectively applied series of feature-value alterations as delineated by a rule chain. As it pertains to the model's response, the metric is immune to confounding issues of feature correlation that might impact the rule's learning process. The fidelity assessment determines whether the model's output corresponds with the predicted class transition upon the application of the entire set of prescribed features' modifications. If one of the altered features exhibits a strong correlation with another feature, the model might derive its information from the latter and not the altered one, potentially preventing any change in prediction as specified by the rule. In such a scenario, the fidelity would invariably be 0, as the rule failed to represent the behavior of the model.

Given that the fidelity metric has been specifically defined for local rules, and RaMiCo is designed to generate global rules, we adopted the localization method provided by RaMiCo in order to extract a collection of local rules applicable to a given sample. Subsequently, these rules were collectively applied to the sample. The parameter for determining the number of rules-records r to be selected each time is user-adjustable within the method. Increasing the parameter r results in a greater number of global rules involved, thereby raising the likelihood of a transition to the target class and, consequently, enhancing the fidelity of the approach. However, this also results in a greater number of rules, thus potentially diminishing its comprehensibility.

To evaluate the fidelity of rules extracted by LORE we followed these steps:

1. Query LORE with each input sample $x \in X$.
2. For each query x, obtain a set of rules S.
3. For each rule $r \in S$ with class transition $c \to p$, verify that both $M(x) = c$ and $M(x) = p$ with $x' \equiv r(x)$.

The InComprehensibility of LORE can be determined directly by applying the provided definition. In contrast, for both RaMiCo and MRT, the active rule chain is contingent upon the localization method employed to associate the sample x with its corresponding set of active rule-records. For RaMiCo, assuming r rule records, InComprehensibility is measured by the number of distinct features present in the r-records. Consequently, the upper bound of InComprehensibility for RaMiCo is r. For MRT, InComprehensibility is calculated as the average depth of the surrogate decision tree as, to achieve sample localization x, it is necessary each time to identify the active leaf node, predict the class

c by applying the tree chain rule from the root to the leaf, and subsequently compute the minimal rule, as elucidated in the preceding section.

For RaMiCo rules, the plausibility of the rules, ρ, defined in Sect. 3, is reported to measure how well the rule reflect the actual data distribution. To facilitate interpretability, the plausibility values were normalized using the minimum and maximum isolation scores of the original training samples. With this normalization, a plausibility score greater than one indicates that the generated sample has an anomaly score greater than any sample in the dataset, while a value below zero implies that the sample has an anomaly score lower than any sample in the dataset. A plausibility value around 0.5, as observed in our experimental results, suggests that the generated samples and the corresponding extracted rules exhibit an anomaly score similar to the average samples in the input space. This provides insight into the extent to which the mined samples align with the natural distribution of the dataset, thereby informing the choice of mining strategy for generating meaningful rule-based explanations.

4.2 Results

4.2.1 White-box rules-extraction experiment: plausibility

In the first experimental setup, we trained and tested a decision tree classifier on all the datasets considered in the study. We then applied the RaMiCo methodology using three different mining strategies–*Random*, *Center*, and *Decision Boundary*–to extract factual and counterfactual samples. For each generated sample, we computed the plausibility value, which quantifies how isolated the sample is within the input space. The boxplots in Fig. 2 present the distribution of plausibility values obtained from different mining strategies, averaged across various hyperparameter configurations of the RaMiCo algorithm, specifically the number of factual and counterfactual samples.

Analyzing the results across the three mining strategies, we observe that the *Center* approach consistently produces the most plausible rules, as indicated by its lower plausibility scores compared to the *Random* and *Decision Boundary* strategies. This outcome is expected, given that the *Center* strategy selects samples that are closer to the core distribution of the dataset, making them less isolated. Conversely, the *Random*

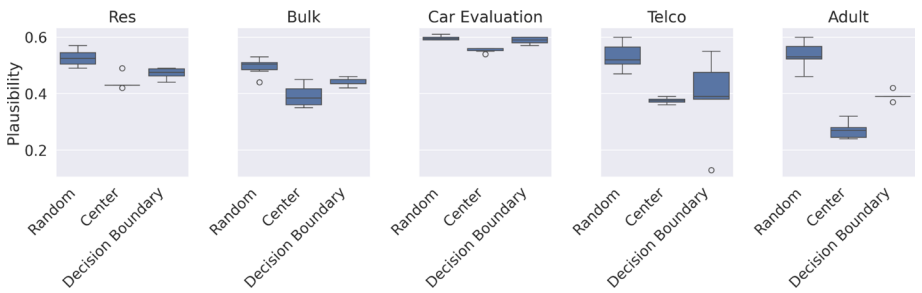

Fig. 2 Comparison of plausibility values for different mining strategies used in the RaMiCo methodology across multiple datasets. The three mining strategies–*Random*, *Center*, and *Decision Boundary*–are employed to extract factual and counterfactual samples from the input space. Plausibility is computed as the isolation score of the sample, normalized by the minimum and maximum isolation scores within each dataset (lower values indicate better plausibility). The boxplots illustrate the distribution of plausibility values for each mining strategy applied to the respective datasets

strategy tends to generate the least plausible rules on average, with higher plausibility values, suggesting that the extracted samples are more isolated from the overall data distribution. This is likely due to the fact that samples are mined from non-congruent space sections, leading to the selection of samples that are less representative of the dataset's general structure.

These results demonstrate the capability of RaMiCo to generate samples and extract rules that remain consistent with the data distribution due to its minimality-driven structure. The sample generation mechanism involves the substitution of one feature at a time, effectively minimizing the introduction of outliers. This controlled modification process ensures that the generated factual and counterfactual samples are plausible within the dataset's original space, reinforcing the reliability of the extracted rules for interpretable model analysis.

Figure 3 shows the percentage of unique rules extracted by RaMiCo under different mining strategies across multiple datasets. A unique rule, is a rule $(c,\ p)$, $a, \rho_a : \mathbb{E}[x_i[a]] \pm \sigma_i[x_i[a]] \rightarrow \mathbb{E}[x_j[a]] \pm \sigma_j[x_j[a]]$ extracted from a RaMiCo mining strategy, which does not appear in the set of the extracted rules of another mining strategy. Two rules $r1$ and $r2$ are considered the same if sharing equal $(c,\ p)$, a and overlapping $\mathbb{E}[x_i[a]] \pm \sigma_i[x_i[a]] \rightarrow \mathbb{E}[x_j[a]] \pm \sigma_j[x_j[a]]$.

Higher percentages indicate that two strategies produce more distinct sets of rules, while lower percentages suggest a larger overlap. Overall, each approach yields more than 30% unique rules, meaning each method extracts at least one-third of its rules that are not found by another method. Adult dataset results in the greatest difference, where 73.1% of the rules extracted by the Random strategy are not identified by the Decision Boundary strategy. This pattern underscores how each approach captures distinct facets of the model's decision-making, making it difficult to single out one method as universally superior. Consequently, RaMiCo combines all these strategies to offer a more exhaustive exploration of the rule space, thereby providing a richer, multi-perspective explanation of the model's behavior.

In the following, we settled RaMiCo to use an aggregation of all of the 3 mining approaches as discussed in the design section.

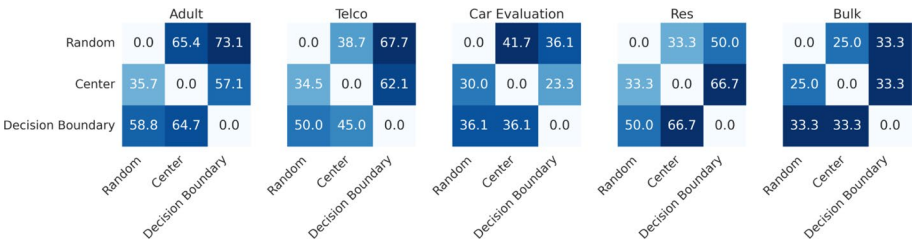

Fig. 3 Comparison of the percentage of unique rules extracted by RaMiCo under different mining strategies (Random, Center, and Decision Boundary) across five datasets (Adult, Telco, Car Evaluation, Res, and Bulk). Each cell shows the percentage of rules unique to the corresponding pair of strategies. Higher values indicate a greater difference between the extracted rule sets, whereas lower values suggest a higher degree of overlap

4.2.2 White-box rules-extraction experiment: coverage

In the second experiment, we trained again a decision tree white-box classifier on the benchmark datasets involved in the study and analyzed how the RaMiCo extracted rules were comparable to the model.

Table 2 describes the number of rules extracted by RaMiCo while varying the number of *factual* and *counterfactual* samples mined . Additionally, the table presents the *Coverage* and *Plausibility* metrics for each configuration.

Table 2 Number of rules extracted from different datasets using the RaMiCo method

Dataset	Factuals	Counterfactuals	Rules	Coverage	Plausibility
Adult	10	50	30.90 ± 4.31	20.61 ± 3.67	0.40 ± 0.02
	25	25	33.10 ± 2.56	26.49 ± 4.34	0.39 ± 0.02
		50	38.30 ± 4.32	30.83 ± 4.55	0.40 ± 0.02
	50	25	38.70 ± 3.59	33.65 ± 5.00	0.40 ± 0.02
		50	44.30 ± 4.37	35.84 ± 5.13	0.40 ± 0.01
	100	50	51.30 ± 5.23	38.87 ± 5.16	0.40 ± 0.02
Bulk	10	50	25.50 ± 5.02	81.00 ± 10.18	0.45 ± 0.04
	25	25	30.80 ± 6.75	89.30 ± 10.16	0.45 ± 0.05
		50	30.80 ± 7.15	89.26 ± 10.14	0.45 ± 0.03
	50	25	34.60 ± 8.38	88.97 ± 10.05	0.45 ± 0.04
		50	35.30 ± 8.49	89.27 ± 10.14	0.45 ± 0.04
	100	50	36.50 ± 8.63	89.86 ± 10.64	0.47 ± 0.04
Car Evaluation	10	50	95.70 ± 7.69	68.08 ± 8.39	0.57 ± 0.02
	25	25	107.80 ± 9.47	74.50 ± 10.17	0.58 ± 0.02
		50	109.40 ± 11.17	75.40 ± 7.88	0.59 ± 0.02
	50	25	115.30 ± 13.74	76.95 ± 6.87	0.60 ± 0.02
		50	116.20 ± 13.77	76.84 ± 6.80	0.60 ± 0.02
	100	50	118.70 ± 13.52	77.33 ± 5.96	0.60 ± 0.02
Res	10	50	24.00 ± 7.82	86.20 ± 8.53	0.44 ± 0.05
	25	25	27.60 ± 9.69	93.05 ± 9.78	0.45 ± 0.03
		50	28.10 ± 9.68	93.05 ± 9.78	0.45 ± 0.03
	50	25	30.90 ± 9.96	93.82 ± 9.34	0.46 ± 0.03
		50	31.90 ± 9.87	93.82 ± 9.34	0.47 ± 0.03
	100	50	32.60 ± 10.23	93.82 ± 9.34	0.47 ± 0.03
Telco	10	50	32.30 ± 5.38	23.34 ± 5.17	0.41 ± 0.04
	25	25	43.00 ± 5.23	33.52 ± 7.14	0.42 ± 0.04
		50	44.20 ± 7.33	36.58 ± 7.62	0.42 ± 0.03
	50	25	56.80 ± 6.91	41.79 ± 6.85	0.42 ± 0.03
		50	55.70 ± 8.46	43.46 ± 7.03	0.41 ± 0.02
	100	50	67.80 ± 8.35	45.96 ± 7.20	0.43 ± 0.03

The method takes as input a specified number of factual and counterfactual samples and extracts rules from a white-box classifier, specifically a decision tree. The *Rules* column indicates the number of rules extracted. *Coverage* represents the percentage of extracted rules that match a leaf node of the decision tree

As observed, most of the results exhibit plausibility values between 0.4 and 0.6, indicating that the plausibility of the generated samples aligns with the dataset distribution. This consistency reinforces the capability of RaMiCo to generate rules that remain within a reasonable anomaly range. Moreover, the number of extracted rules increases as the number of *factual* and *counterfactual* samples grows, demonstrating a direct relationship between the amount of the RaMiCo input data and the completeness of the mined rules.

A similar trend is evident in the *Coverage* values, which measure the proportion of extracted rules that align with the decision tree structure. As the number of factual and counterfactual samples increases, RaMiCo achieves higher tree coverage, indicating that a greater portion of the model's learned decision logic is successfully captured. This trend is consistent across all datasets. In most cases, RaMiCo is capable of mining more than 70% of the decision tree structure, highlighting its effectiveness in extracting meaningful and representative rules.

For the Adult, Telco, Bulk, and Res datasets, an increase in the number of factuals (and correspondingly in the number of counterfactuals) leads to a consistent increase in the number of extracted rules and in the tree coverage. For instance, in the Adult dataset, the number of rules grows from 11 to 18, while the coverage increases from 25.00 to 40.40%. Similarly, the Telco dataset shows an increase from 8 to 17 rules with coverage rising from 14.10% to 29.32%. This trend indicates that a larger set of input samples (factuals and counterfactuals) enables the extraction of more comprehensive rule sets that better represent the decision tree structure, thereby capturing a greater portion of the tree's logic.

With the Car Evaluation dataset, the number of extracted rules remains stable between 44 and 51, with coverage consistently around 67.85% to 75.00%, due to its fully categorical nature with 6 features and about 4 distinct categories each. In these datasets, minimal feature transitions are limited. This can be explained by considering the presence of different categorical feature. For each categorical feature, transitions involve changing one category to another, leading to about 4 choices for the original category and 3 alternatives, resulting in 12 potential transitions per feature. With 6 features, this results in around $6 \times 12 = 72$ distinct minimal rules. Thus, the stabilization in extracted rules likely reflects the limited combinatorial possibilities in a categorical dataset, quickly exhausting viable transitions.

4.2.3 Black-box rules-extraction experiment: fidelity

To evaluate the fidelity of RaMiCo in explaining black-box models, we trained two machine learning classifiers, a Multi-Layer Perceptron (MLP) and an XGBoost model, on different benchmark datasets. After training, we applied RaMiCo and two state-of-the-art rule extraction methods, LORE and MRT, to generate rules that approximate the model decision-making process. We then measured the fidelity of the extracted rules on the training set, which quantifies how accurately they replicate the predictions of the black-box model.

Table 3 presents the fidelity results across different datasets and classifiers. Higher fidelity values indicate that the extracted rules align more closely with the black-box model's decisions, values highlighted in bold indicate the best method for dataset. To assess the statistical significance of the observed differences with RaMiCo results, we performed the paired Wilcoxon signed-rank tests for each configuration (RaMiCo vs LORE and RaMiCo vs MRT), considering 10 repetitions of the same experiment. $p - values$ greater than 0.05, i.e. no statistical significance, are marked with an asterisk (*) in the table. As observed, RaMiCo consistently outperforms LORE across all datasets and model configurations,

Table 3 Fidelity results of the proposed approach (RaMiCo) compared with two state-of-the-art methods (LORE and MRT) on various datasets

Model	Method	Adult (%)	Bulk (%)	Car evaluation (%)	Res (%)	Telco (%)
MLP	RaMiCo	**67.14**	**83.41**	80.91	**94.33**	**45.24**
	LORE	56.00	61.60	69.60	77.20	36.56
	MRT	54.00	73.72	**91.83**	84.13	42.79
XGBoost	RaMiCo	**82.23**	70.57	80.79	82.51	**49.51**
	LORE	70.35	58.50	67.15	72.30	44.09*
	MRT	59.45	**73.64**	**91.83**	**87.54**	42.58

The fidelity metric measures how closely the rules extracted by black-box classifiers (MLP and XGBoost) resemble the model's behaviour. Asterisk (*) on the Fidelity value of LORE or MRT indicates statistical significance (p-values) of the paired (RaMiCo vs LORE or RaMiCo vs MRT) Wilcoxon test greater than 0.05 across 10 repetitions

demonstrating its effectiveness in capturing the decision logic of the underlying models. Only in the Telco-XGboost configuration, RaMiCo and LORE are not statistically different, i.e. the p-value is higher than 0.05. RaMiCo achieves the highest fidelity for MLP on the Adult, Bulk, and Res datasets, highlighting its capability to provide faithful and interpretable rule sets. However, in some cases, particularly on the Car Evaluation datasets, MRT yields the highest fidelity.

Across the different datasets and models, RaMiCo reached 73.7% average fidelity, while LORE and MRT got 61.3% and 70.15%. The mean difference in fidelity between RaMiCo and LORE across all datasets and models is 12.33%, demonstrating a significant advantage of RaMiCo in extracting rules that better resemble the black-box model's decision-making. The largest improvement over LORE is observed in the Bulk dataset, where RaMiCo achieves by average 77% fidelity compared to LORE's 60.05%, yielding a difference of 16.94%. Conversely, the smallest improvement is in the Telco dataset, where RaMiCo surpasses LORE by 7.05% (47.38% vs. 40.32%).

When comparing RaMiCo with MRT, the mean difference in fidelity is 3.51%, indicating that MRT performs better than LORE but still lags behind RaMiCo. The largest difference, favoring MRT, is in the Car Evaluation dataset, where RaMiCo achieves 80.85% fidelity, while MRT reaches 91.83%, resulting in a difference of −10.99%. The largest difference favouring RaMiCo instead is in the Adult dataset, where RaMiCo gets 74.69% and MRT gets 56.72%, resulting in a difference of 17.96%.

RaMiCo demonstrates superior performance compared to MRT for both the XGBoost and MLP models in the Adult and Telco datasets. Conversely, MRT exhibits better performance than RaMiCo for both models in the Car Evaluation dataset. These findings indicate that RaMiCo and MRT generally exhibit comparable performance levels, with RaMiCo achieving higher fidelity on average.

Figure 4 shows RaMiCo's fidelity improving as more rules-records r are used in the localisation mechanism. As r increases, the fidelity metric on the y-axis rises for both MLP and XGBoost models, enhancing the alignment between RaMiCo's rules and the black-box model's decisions. This is particularly evident in XGBoost, which maintains higher fidelity across all r values than MLP.

Within the localisation mechanism (i.e. to generate local rules) where a single feature is present in all k-rule records, RaMiCo generates a single, generalized rule by

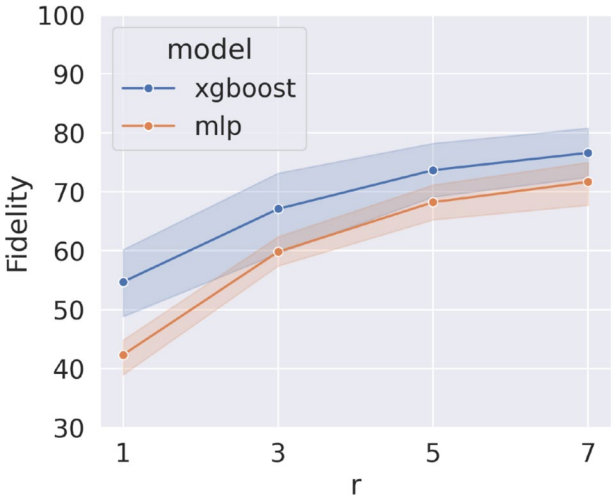

Fig. 4 Fidelity of RaMiCo in explaining two black-box classifiers, MLP and XGBoost, trained on the Adult dataset. The fidelity measures how well the extracted rules approximate the predictions of the original model. The parameter r represents the number of rule-records involved in the localisation mechanism of the RaMiCo global rules. As r increases, fidelity improves for both models. The shaded regions indicate variability across multiple parameter setups (number of factuals and counterfactuals) of the RaMiCo method

employing their mean value. In contrast, for records containing multiple features in k-rules, the rule integrates multiple conditions, thereby establishing a chained decision logic. Consequently, this approach may result in rules that are less minimal or harder to comprehend.

In the final analysis, we assess the interpretability of the extracted rules by comparing the InComprehensibility metric, defined as the average rule length, across the different rule extraction methods. The rule length corresponds to the number of logical conditions that constitute the rule's logic chain; thus, longer rules indicate a less minimal and more complex explanation, which in turn renders them more difficult for users to comprehend.

Figure 5 presents the average rule length for each method–RaMiCo, LORE, and MRT with different datasets. As observed, RaMiCo consistently yields shorter rules, suggesting that its counterfactual extraction process effectively identifies minimal modifications necessary for model explanation. In contrast, the MRT approach frequently produces longer

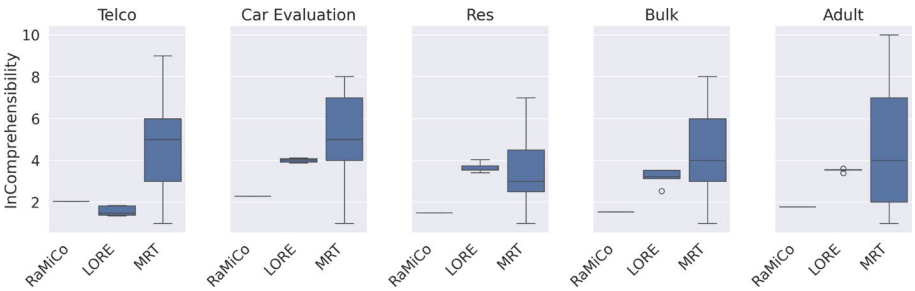

Fig. 5 Comparison of the InComprehensibility (average rule length) across different rule extraction methods–RaMiCo, LORE, and MRT on multiple datasets (Telco, Car Evaluation, Res, Bulk, and Adult). The rule length is defined as the number of logical conditions in each extracted rule. Higher values indicate more complex and less interpretable rules. RaMiCo consistently produces the shortest rules, demonstrating its effectiveness in generating minimal and comprehensible explanations. In contrast, MRT often results in longer rules, leading to higher incomprehensibility

rule chains, indicating a higher level of incomprehensibility. LORE, while generally more concise than MRT, exhibits intermediate rule lengths relative to the other methods.

These results demonstrate that, in addition to providing high fidelity and robust coverage, RaMiCo excels in generating interpretable rules. The concise nature of the rules extracted by RaMiCo facilitates a clearer understanding of the underlying decision logic.

5 Conclusion

In this study, we present RaMiCo (Region Aware Minimal Counterfactual Rules), an innovative methodology designed for the extraction of minimal yet highly informative counterfactual rules from black-box classifiers. Our experimental results across different datasets and experimental configurations have proved that: (1) RaMiCo demonstrates elevated fidelity in replicating the decision-making processes of black-box models; (2) by manipulating the localization parameter k, users can effectively navigate the trade-off between rule minimality (i.e., comprehensibility) and fidelity; (3) when evaluated with white-box classifiers, such as decision trees, RaMiCo exhibits substantial congruence with their intrinsic structures; (4) in comparison to different state-of-the-art methodologies (e.g. MRT and LORE) for rule extraction, RaMiCo exhibits superior performance in terms of both fidelity and comprehensibility, indicating that RaMiCo encapsulates the model behavior with a minimal set of rules.

By concentrating on single-feature substitutions, RaMiCo extracts rules that are both minimal and region-aware, thereby generating explanations that closely reflect the inherent distribution of the data. This minimality is demonstrated by the lower InComprehensibility scores observed, as RaMiCo consistently produces shorter rule chains compared to competing methods. Moreover, the approach exhibits a clear scaling behavior: as the number of factual and counterfactual samples increases, both the number of extracted rules and their coverage of the decision tree also rise, resulting in a more comprehensive representation of the classifier's internal logic. However, this trend is influenced by the intrinsic characteristics of the dataset; for instance, in fully categorical datasets such as Car Evaluation, the limited variety of unique transitions leads to a stabilization in the number of extractable rules.

However, while the minimality of the generated rules enhances interpretability, it may sometimes result in oversimplified explanations that fail to capture complex interactions among multiple features. In these cases, methods that employ more intricate rule logic chains can achieve higher fidelity, as evidenced in Table 3, where MRT, despite producing less comprehensible rules, occasionally outperforms RaMiCo.

RaMiCo's design presupposes that each input feature corresponds to a semantically meaningful attribute whose perturbation yields an intelligible change in model output. Consequently, its direct application to unstructured, high-dimensional domains–such as raw image pixels or deep latent embeddings–remains limited: flipping an individual pixel or embedding coordinate typically produces neither coherent nor interpretable counterfactuals, and mining factually and counterfactually relevant instances in such spaces is both computationally prohibitive and conceptually ambiguous. To extend RaMiCo beyond tabular or otherwise intrinsically interpretable data, one promising direction is to perform instance mining and rule extraction in a lower-dimensional, conceptually grounded latent space (e.g., via clustering or prototype selection in a hidden-layer representation). In this manner, each latent "feature" may capture a higher-level construct amenable to meaningful

counterfactual transitions, thereby preserving RaMiCo's rule-based explanatory paradigm while mitigating the challenges of raw high-dimensional inputs.

In conclusion, RaMiCo constitutes a significant advancement in interpretable machine learning by offering a model-agnostic approach to extract counterfactual rule explanations. The rules generated by RaMiCo provide a global perspective on the decision-making process of the black-box model, while also enabling the localization of these global rules to specific input instances. Its ability to generate high-fidelity, minimally complex rules positions it as a valuable resource for applications where trust and interpretability are critical. The results presented in this study underscore the potential of RaMiCo to enhance our understanding of complex predictive models and support more informed decision-making in high-stakes environments.

An additional consideration concerns RaMiCo's current limitations and possible future developments. RaMiCo generates minimal rules that involve changing only one feature at a time. Although this design ensures highly interpretable explanations, it may prove insufficient for complex models operating on large sets of features, such as neural networks classifying raw, unstructured data (e.g., images). In such cases, modifying a single input feature may not elicit a meaningful change in the model's behavior. A promising avenue for future work involves grouping semantically related features, allowing simultaneous modifications across multiple dimensions. This improvement could better capture complex decision-making processes of high-dimensional models while preserving interpretability as a central goal.

Acknowledgements Work partially supported by: (i) the European Commission under the NextGenerationEU program, Extended Partnership PNRR PE1—"FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-centered AI", and PNRR—M4 C2, Investment 1.5 "Creating and strengthening of "innovation ecosystems", building "territorial R&D leaders", project "THE—Tuscany Health Ecosystem", Spoke 6 "Precision Medicine and Personalized Healthcare"; (ii) the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence), in the framework of the "Reasoning" project, PRIN 2020 LS Programme, Project Number 2493 04-11-2021, and in the framework of the project "OCAX -Oral CAncer eXplained by DL-enhanced case-based classification" PRIN 2022 code P2022KMWX3; (iii) the Italian Ministry of Enterprises and Made in Italy, in the framework of the "Agreements for Innovation" Project "4DDS—4D Drone Swarms" Ref. no. F/310097/01-04/X56. (iv) the Tuscany Region in the framework of the WAU project, PR FESR 2021–2027, Project No. 27716.29122023.042000115; (v) the Digital Republic Fund supported by Google.org in the framework of the wAIne project, CrescerAI national call, Project No. 2023-CRE-00288, ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the explanation of AI decision making* (https://xai-project.eu/index.html), the NextGenerationEU programme under the funding scheme "*SoBigData.it*—Strengthening the Italian RI for Social Mining and Big Data Analytics"—Prot. IR0000013, and by the Italian Project Fondo Italiano per la Scienza FIS00001966 MIMOSA. Financial support was provided by the Research Foundation Flanders (FWO) via SBO mandate 1SH4Z24N. The authors thank Claudio Daka, a master's student at the University of Pisa, for his support in this research as part of his Master Thesis.

Author contributions Guido Gagliardi: conceptualization, formal analysis, investigation, software, visualization, writing—original draft. Antonio Luca Alfeo: conceptualization, formal analysis, investigation, software, writing—review & editing. Riccardo Guidotti: formal analysis, investigation, funding acquisition, writing—review & editing. Mario G. C. A. Cimino: formal analysis, visualization, funding acquisition, writing—review & editing.

Data availability The datasets used in this study, i.e. Adult, Telco Costumer Churn, Car Evaluation, are publicly available at the UCI Machine Learning Repository as referenced in the paper. Due to confidentiality agreements, the Res and Bulk datasets can only be made available to bona fide researchers subject to a non-disclosure agreement.

Code availability The code of the proposed approach will be published on GitHub once the paper has been accepted.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Alfeo, A. L., Cimino, M. G., & Gagliardi, G. (2023). Concept-wise granular computing for explainable artificial intelligence. *Granular Computing, 8*(4), 827–838.

Alfeo, A. L., Zippo, A. G., Catrambone, V., et al. (2023). From local counterfactuals to global feature importance: Efficient, robust, and model-agnostic explanations for brain connectivity networks. *Computer Methods and Programs in Biomedicine, 236*, 107550.

Amoukou, S. I., & Brunel, N. J. (2022). Rethinking counterfactual explanations as local and regional counterfactual policies. arXiv preprint arXiv:2209.14568

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion, 58*, 82–115.

Awal, M. A., & Roy, C. K. (2024). Evaluatexai: A framework to evaluate the reliability and consistency of rule-based xai techniques for software analytics tasks. *Journal of Systems and Software, 217*, 112159.

Bénard, C., Biau, G., Da Veiga, S., et al. (2021). Sirus: Stable and interpretable rule set for classification. *Electronic Journal of Statistics, 15*, 427–505.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology, 67*(1), 135–157.

Chakraborty, M., Biswas, S. K., & Purkayastha, B. (2020). Rule extraction from neural network trained using deep belief network and back propagation. *Knowledge and Information Systems, 62*(9), 3753–3781.

Chen, Y., Cui, T., Capstick, A., Fletcher-Loyd, N., & Barnaghi, P. (2024). Enabling regional explainability by automatic and model-agnostic rule extraction. arXiv preprint arXiv:2406.17885

Chou, Y. L., Moreira, C., Bruza, P., et al. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion, 81*, 59–83.

De Fortuny, E. J., & Martens, D. (2015). Active learning-based pedagogical rule extraction. *IEEE Transactions on Neural Networks and Learning Systems, 26*(11), 2664–2677.

Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics, 7*(4), 277–287.

Elkhawaga, G., Abu-Elkheir, M., & Reichert, M. (2023). A rule-based evaluation method of local explainers for predictive process monitoring. In *2023 IEEE international conference on data mining workshops (ICDMW)* (pp. 922–930). IEEE.

Foulsham, M., Hitchen, B., & Denley, A. (2019). *GDPR: How to achieve and maintain compliance*. Routledge.

Freiesleben, T., & König, G. (2023). Dear xai community, we need to talk! fundamental misconceptions in current xai research. In *World conference on explainable artificial intelligence* (pp. 48–65). Springer.

Gagliardi, G., Alfeo A. L., Catrambone, V., Cimino, M. G., De Vos, M., & Valenzal, G. (2023). Fine-grained emotion recognition using brain-heart interplay measurements and explainable convolutional neural networks. In *2023 11th international IEEE/EMBS conference on neural engineering (NER)* (pp. 1–6). IEEE.

Guidotti, R. (2024). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery, 38*(5), 2770–2824.

Guidotti, R., Monreale, A., Ruggieri, S., et al. (2024). Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Mining and Knowledge Discovery, 38*(5), 2825–2862.

Hao, J., Luo, S., & Pan, L. (2022). Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes. *Scientific Reports, 12*(1), 9858.

Ley, Dan, Saumitra Mishra, and Daniele Magazzeni. "Global counterfactual explanations: Investigations, implementations and improvements." arXiv preprint arXiv:2204.06917 (2022).

Love, P. E., Fang, W., Matthews, J., et al. (2023). Explainable artificial intelligence (xai): Precepts, models, and opportunities for research in construction. *Advanced Engineering Informatics, 57*, 102024.

Luo, G., Johnson, M. D., Nkoy, F. L., et al. (2020). Automatically explaining machine learning prediction results on asthma hospital visits in patients with asthma: Secondary analysis. *JMIR Medical Informatics, 8*(12), e21965.
Mahya, P., & Fürnkranz, J. (2023). An empirical comparison of interpretable models to post-hoc explanations. *AI, 4*(2), 426–436.
Marshakov, D. (2021). Rule extraction from the artificial neural network. In *IOP conference series: Materials science and engineering* (p. 012127). IOP Publishing.
Pornprasit, C., Tantithamthavorn, C., Jiarpakdee, J., Fu, M., & Thongtanunam, P. (2021). Pyexplainer: Explaining the predictions of just-in-time defect models. In *2021 36th IEEE/ACM international conference on automated software engineering (ASE)* (pp. 407–418). IEEE.
Rapp, M., Fürnkranz, J., & Hüllermeier, E. (2024). On the efficient implementation of classification rule learning. *Advances in Data Analysis and Classification, 18*(4), 851–892.
Rawal, K., & Lakkaraju, H. (2020). Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems, 33*, 12187–12198.
Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems, 7*(1), 2400304.
Schoenborn, J. M., & Althoff, K. D. (2019). Recent trends in xai: A broad overview on current approaches, methodologies and interactions. In *ICCBR workshops* (pp. 51–60).
van der Waa, J., Nieuwburg, E., Cremers, A., et al. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, 103404.
Zarlenga, M. E., Shams, Z., & Jamnik, M. (2021). Efficient decompositional rule extraction for deep neural networks. arXiv preprint arXiv:2111.12628

Authors and Affiliations

Guido Gagliardi[1,3,4] · Antonio Luca Alfeo[7] · Riccardo Guidotti[5,6] · Mario G. C. A. Cimino[1,2]

✉ Guido Gagliardi
guido.gagliardi@phd.unipi.it

Antonio Luca Alfeo
antonioluca.alfeo@uniecampus.it

Riccardo Guidotti
riccardo.guidotti@unipi.it

Mario G. C. A. Cimino
mario.cimino@unipi.it

[1] Department of Information Engineering, University of Pisa, Pisa, Italy

[2] Research Center E. Piaggio, University of Pisa, Pisa, Italy

[3] Department of Information Engineering, University of Florence, Florence, Italy

[4] Department of Electrical Engineering, KU Leuven, Leuven, Belgium

[5] Department of Computer Science, University of Pisa, Pisa, Italy

[6] ISTI, CNR, Pisa, Italy

[7] Department of Theoretical and Applied Sciences, eCampus University, Novedrate, Italy