

ORIGINAL RESEARCH

Open access

Comparing Intrinsic Interpretability and Post-Hoc Explanations: An Empirical Study across Multiple Datasets

Pierre Dubois^{1*}, Marc Lefevre², Claire Moreau¹, Julien Martin³, Thomas Dupont⁴

Abstract

In recent years, considerable attention has been directed toward clarifying complex, opaque systems, commonly referred to as black-box models, including deep neural networks and random forests. So-called model-agnostic techniques generally approximate the outputs of these non-transparent systems using an interpretable architecture, ignoring the internal mechanisms of the primary black-box system itself. Consequently, it remains an important inquiry whether the straightforward training of transparent, white-box architectures ought to be favored over the post-hoc approximation of these opaque models. This study presents the findings of an empirical investigation comparing post-hoc explanatory methods against inherently interpretable architectures across multiple datasets, focusing on both rule-based and feature-based frameworks. The empirical outcomes indicate that directly trained transparent models frequently approximate black-box models with an accuracy that matches or exceeds that of post-hoc surrogate models, despite lacking direct access to the underlying black-box structure.

Keywords Explainable AI, Interpretable machine learning, Interpretable models, Black-box explanation, White-box models

*Correspondence:

Pierre Dubois

pierre.dubois@u-bordeaux.fr

¹ Department of Intrinsic Interpretability and XAI Comparison, Faculty of Engineering, University of Bordeaux, Bordeaux, France

² Department of Post-Hoc Explanations and Empirical Studies, Faculty of Engineering, University of Nantes, Nantes, France

³ Department of Multi-Dataset Explainability Evaluation, Faculty of Engineering, University of Strasbourg, Strasbourg, France

⁴ Department of Interpretable vs. Post-Hoc Models, Faculty of Engineering, Université Paris-Saclay, Paris, France

Introduction

Machine learning techniques have been broadly deployed across a variety of fields and use cases, including healthcare and financial services. Frequently, the resulting architectures function as black-box models, meaning their internal representations lack straightforward interpretability. Consequently, the predictions generated by these systems are difficult for humans to comprehend.

The critical requirement for explanations to clarify the operational mechanics of these systems has driven extensive research into understanding trained models [1]. These approaches can be categorized into local explanations, which attempt to approximate the black-box model in the immediate neighborhood of a specific instance being analyzed (e.g., [2, 3]), and global models, which aim to represent the overall behavior of the entire black-box system through an interpretable surrogate. Recently, numerous frameworks have explored the synthesis of global models from sets of local explanations (e.g., [4, 5]). Additionally, a distinction can be made between model-specific explanation techniques designed for particular classes of black-box architectures like deep neural networks (e.g., [6]), and model-agnostic explanation techniques that operate independently of the underlying black-box architecture's characteristics (e.g., [2]).

Although the necessity of explaining black-box systems is undeniable in high-stakes decision-making scenarios, multiple difficulties and complications have revitalized interest in training inherently interpretable architectures, such as decision trees or rule induction systems, from the outset.

The primary limitation is that post-hoc explanatory techniques merely approximate the primary black-box system, meaning the generated explanations frequently fail to precisely map the true behavior of the target model. This phenomenon is commonly monitored by evaluating the fidelity of the surrogate system, which measures how closely it reproduces the target model's decisions. Furthermore, even in an ideal scenario where an explanation operates without error, it might utilize entirely distinct features, indicating that the explanation lacks fidelity to the actual computational path of the black-box model. Additionally, if the underlying black-box model contains flaws, debugging becomes highly complex because both the explanatory framework and the black-box architecture must be simultaneously debugged and maintained. Given these limitations, researchers have argued that greater emphasis should be placed on developing highly accurate, inherently interpretable models [7].

Prompted by these considerations, this study assesses the degree to which post-hoc explanations can be substituted by directly trained interpretable techniques that operate without knowledge of the underlying black-box systems. The objective is to determine if the predictive performance of an interpretable model is sufficiently robust to replace model-agnostic frameworks, or conversely, to quantify the information loss associated with such a substitution. To achieve this, the effectiveness of both local and global explanation techniques is examined empirically to evaluate the validity of this premise. We execute a sequence of experiments to analyze and compare the performance of multiple interpretable frameworks against black-box explanations. The evidence gathered from rule-based and feature-based explanatory frameworks appears to support our core hypothesis.

Related Work

A substantial body of literature has explored the explainability and interpretability of black-box models. We point to [8-10] for comprehensive overviews and restrict our discussion to a brief summary of the literature most relevant to our investigation. Specifically, we focus on contrasting studies within two primary streams: the direct training of interpretable systems and post-hoc explanation frameworks.

Direct learning of interpretable models

Directly trained interpretable techniques comprise a category of algorithms that construct transparent models without relying on a primary black-box architecture. Linear regression, logistic regression, and decision trees represent the most conventional transparent frameworks. Our primary focus centers on the induction of logical rules [11], an area where Ripper (Repeated Incremental Pruning to Produce Error Reduction) continues to serve as a highly competitive, state-of-the-art benchmark [12]. In the domain of feature-based explanations, we utilize GA²Ms as an interpretable (white-box) framework derived from Generalized Additive Models. This approach is engineered to achieve competitive predictive performance relative to advanced machine learning models while preserving structural intelligibility and explainability [13].

Local explanations of black-box models

Post-hoc explanation systems can either deliver a global overview of the entire black-box model (global surrogates) or provide a localized explanation targeted at a specific instance (local surrogates). Several prominent approaches are model-agnostic, meaning they are compatible with any underlying black-box architecture.

Shap [3] is a framework grounded in game theory that explains predictions by calculating post-hoc weights that quantify the significance of an input feature toward the final output. Ribeiro et al. [2] introduced Local Interpretable Model-agnostic Explanations (LIME), which centers on clarifying the specific predictions of a black-box model. While LIME is theoretically independent of the specific interpretable model selected for the explanation, it is predominantly deployed to generate

feature-weight-based explanations. Local Rule-based Explanations (LORE) represents a variation of LIME that is specifically configured to output rules as local surrogate models [14]. It delivers interpretable and locally faithful explanations by training a local interpretable model on a synthetic neighborhood produced via a genetic algorithm. This pipeline then extracts structured rules from the local estimator to form a coherent explanation. Model Agnostic Supervised Local Explanations (MAPLE) operates as a hybrid framework capable of functioning both as a highly precise tree-based predictive model and as a feature-based local explanatory tool [15].

Combining local explanations into global models

While interpretable architectures generally operate globally—meaning they offer a mechanism to classify any arbitrary point within the data space—post-hoc explanations are typically local, meaning they apply exclusively to the specific instance for which they were generated. Rule-based frameworks are notably compelling in this domain because a global model typically manifests as a rule set or rule list built from discrete rules, which can themselves be conceptualized as local models [16, 17]. More broadly, a number of recent initiatives have concentrated on integrating local explanations into comprehensive global models, which simplifies the comparison between local and global systems.

For instance, Yang *et al.* [18] introduced Global Interpretation via Recursive Partitioning (GIRP), a framework designed to construct a global interpretation tree for diverse machine learning models using their local explanations. This approach sequentially divides the input variable space by optimizing the average contribution of input features derived from local explanations across the separated sub-spaces. The final output of this process is a binary interpretation tree that maps out a collection of decision rules approximating the target model. van der Linden *et al.* [19] presented Global Aggregations of Local Explanations (GALE) as a strategy to illuminate the global decision-making pathway. Their work evaluates the degree to which local explanations can provide comprehensive, macro-level insights into a black-box model. To accomplish this, local explanations generated by the LIME framework are examined and synthesized using multiple aggregation strategies to test their capacity to represent global behavior. Lastly, Setzu *et al.* [4] developed the GLocal to loCAL explainer (GLocalX), which functions as a "local-first" model-agnostic explanation framework. This methodology seeks to exploit local explanation tools and their inherent advantages to construct a global overview. The pipeline initiates from local decision rules and progressively builds global explanations through systematic aggregation.

Methods and Experimental Setup

Problem statement

This investigation addresses the empirical validation of the premise presented by Rudin [7], which argues that artificial intelligence research should prioritize the development of inherently transparent systems rather than the post-hoc interpretation of black-box architectures. To explore this, we select and systematically contrast pairs of model-agnostic post-hoc explanation frameworks and independent, directly trained interpretable algorithms that generate identical syntactic classes of models. Specifically, as illustrated in Figure 1, a black-box system $\mathcal{M}$ is optimized using a designated training dataset. This training set contains n individual instances $\{(\mathbf{x}_i, y_i) | i=1 \cdots n\}$, where each instance is characterized by m distinct features and a corresponding target label $y_i \in Y$.

Subsequently, standard explainable AI methodologies are applied to approximate the decision boundaries of $\mathcal{M}$ via an interpretable surrogate system $\mathcal{J}\mathcal{M}$. Concurrently, a syntactically comparable transparent model $\mathcal{J}$ is trained directly from the source data to serve as a baseline for comparison against $\mathcal{J}\mathcal{M}$.

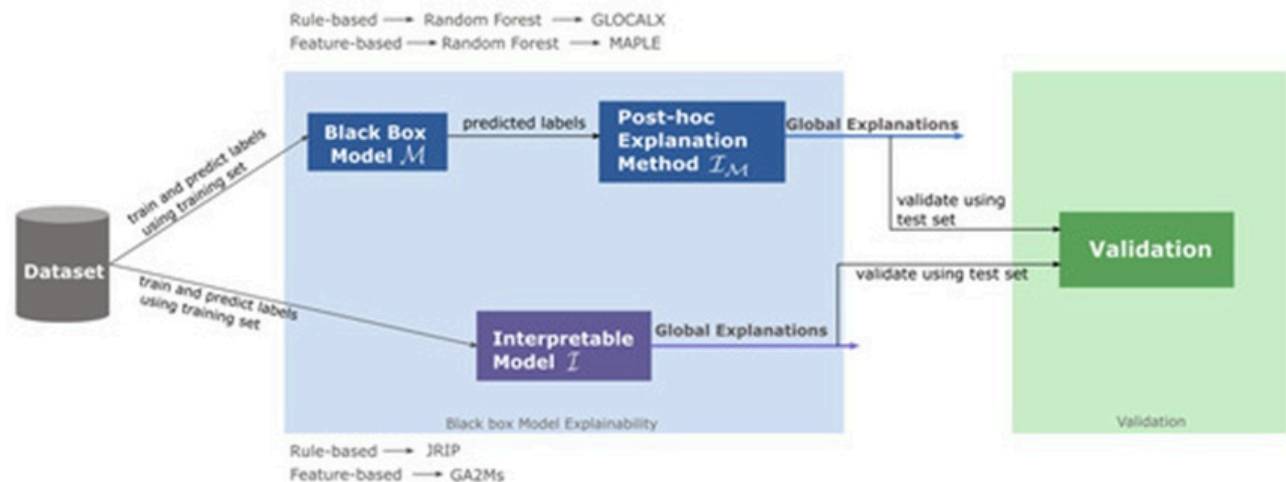


Figure 1. Experimental setup for comparing post-hoc explanations to directly trained interpretable models.

Consequently, the core research question investigated here is the extent to which an inherently interpretable model $\mathcal{I}$, learned directly from data, can successfully approximate an independently optimized black-box system $\mathcal{M}$, and to what degree fidelity degrades relative to an interpretable surrogate $\mathcal{IM}$ that possessed direct access to $\mathcal{M}$. Theoretically, one anticipates that $\mathcal{IM}$ will exhibit superior fidelity (and potentially higher predictive accuracy) compared to $\mathcal{I}$, given that $\mathcal{IM}$ explicitly fits the outputs of $\mathcal{M}$, whereas $\mathcal{I}$ is optimized independently. Nevertheless, because both architectures are trained on the same data distributions, implicit correlations may materialize between them.

Furthermore, it is widely acknowledged that transparent models $\mathcal{I}$ often display lower raw accuracy than complex systems $\mathcal{M}$ because they fundamentally function as approximations of the underlying black-box model $\mathcal{M}$. This approximation quality is commonly quantified via fidelity, which measures the degree to which $\mathcal{IM}$ accurately mirrors the predictions generated by $\mathcal{M}$.

Thus, our objective is to analyze how the two architectures, $\mathcal{I}$ and $\mathcal{IM}$, compare across standard performance metrics—such as structural complexity and predictive accuracy—as well as across these specific fidelity measures.

Additionally, multiple distinct explanations can potentially account for a model's behavior due to the Rashomon effect [20]. This concept dictates that for structured representations like decision trees or rule lists, there frequently exist multiple distinct model configurations that fit the empirical data with equal efficacy. We aim to determine whether substantive differences emerge among the explanations yielded by our selected interpretation methodologies for a given model.

In general, our focus centers on rule-based and feature-based frameworks. We evaluate rule-based methods by examining the logical rule sets they generate, and feature-based methods by analyzing the feature weights assigned to individual variables. The subsequent sections outline the specific algorithms and techniques chosen for this study.

Rule-based interpretability methods

For our rule-based evaluation, GLocalX and JRip are selected to represent model-agnostic and inherently interpretable frameworks, respectively. Both techniques produce explanations formatted as rule sets, which are preferred due to their compact representation and strong alignment with human cognitive reasoning [11].

JRip [12] is a traditional inductive rule learner that constructs rule sets through a sequence of three core operations: growing, pruning, and optimization. Prior to generating each individual rule, JRip partitions the available training instances into two distinct subsets: a growing set utilized to formulate the initial rule, and a pruning set utilized to simplify and generalize the newly learned rule. The complete rule set undergoes further optimization by iteratively re-learning

individual rules within the context of the wider rule base once an adequate volume of positive instances has been successfully accounted for.

GLocalX [4] constructs comprehensive global explanations for an opaque system by aggregating local explanations produced by a local surrogate architecture, such as LORE [14], alongside the predicted target labels from the black-box model. The framework accepts a collection of local explanations as its input and iteratively merges them to establish broader, more generalized logical rules. During each iteration, the local explanations are arranged in a priority queue based on structural similarity, and a batch of data is sampled to evaluate potential merges between candidate explanations. A merge operation is executed when a pair of highly similar explanations is extracted from the queue. This function relies on specific *cut* and *join* operators, enabling the algorithm to generalize explanations while balancing the trade-off between fidelity and model complexity.

To fuse two local explanations, E_i and E_j , the *join* operator is applied to non-conflicting rules, whereas the *cut* operator resolves conflicting rules. Consequently, *join* expands rule generality at the expense of local fidelity, while *cut* specializes the rules to preserve fidelity at the expense of generality. If the resulting merged rule satisfies predefined constraints for simplicity and predictive accuracy, the merged pair is retained, and the original expressions E_i and E_j are replaced. Ultimately, explanations exhibiting suboptimal fidelity are eliminated using the α parameter, which defines a per-class trimming threshold.

Feature-based interpretability methods

Within the domain of feature-based interpretability, GA²Ms [13] is selected as an inherently transparent glass-box algorithm, while MAPLE is chosen to represent post-hoc explanatory frameworks. Both GA²Ms and MAPLE rely on underlying linear formulations, producing feature weights that quantify the relative contribution of each variable to the final prediction.

The GA²Ms algorithm builds upon Generalized Additive Models (GAM), which represent an extension of generalized linear modeling. Rather than relying on simple linear weights, a GAM assumes that the target function can be modeled as a sum of arbitrary shape functions. GA²Ms expands upon standard GAM formulations by incorporating explicit terms that capture pairwise interactions between distinct feature values:

$$g(E(Y)) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j) \quad (1)$$

The training pipeline begins by constructing a shallow decision tree for each input feature independently using a boosting framework, ensuring each individual tree isolates only one feature. This process is repeated across a set number of iterations, eventually producing an ensemble of trees for every variable. Next, the ensembles generated for each feature are transformed into a continuous graph by plotting the predictions of each tree. This results in a final model composed of one graph per feature. Because GA²Ms retains an additive structure, the explicit contribution of each feature to the overall prediction can be straightforwardly interpreted [13].

For inference within the GA²Ms framework, each individual shape function f_i maps a feature value to a specific term contribution, acting much like a lookup table. The retrieved term contributions are summed, and the final predictions are computed by passing this total through a link function g . This additive architecture allows GA²Ms to isolate and report the precise impact of each input feature on the resulting output.

MAPLE combines classical linear regression with a tree-based interpretation of tree ensembles, utilizing a supervised neighborhood approach and feature selection to produce both global and instance-level explanations. The algorithm initially identifies the specific training points within the dataset that are most relevant to a given prediction. It then

computes similarity weights for each training instance x_i by determining how frequently x_i and a target instance x are assigned to the same terminal leaf node across an ensemble of trees $\mathbf{T}=\{T_1, T_2, \dots, T_k\}$, as formalized in [Eq. 2](#):

$$w_i = \frac{1}{k} \sum_{j=1}^k \mathbb{I}[T_j(x_i) = T_j(x)] \quad (2)$$

The calculated weights of the training instances are subsequently embedded into a local linear model to generate both a prediction and a local explanation by solving the weighted linear regression problem outlined in:

$$f_{\text{MAPLE}(x)} = \widehat{\beta}_x^T X \quad \text{where} \quad \widehat{\beta}_x = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n w_i (\beta^T x_i - y_i)^2. \quad (3)$$

Experimental setup

Our empirical evaluation was performed across several standard benchmark datasets, primarily retrieved from the UCI Machine Learning Repository [\[21\]](#). All selected datasets represent binary classification tasks:

- **Adult:** The objective is to determine whether an individual's annual income exceeds 50K.
- **COMPAS Two-Year:** Contains recidivism risk scores predicting the likelihood of an individual reoffending within a two-year window.
- **German Credit:** Identifies whether a loan applicant presents a good or bad credit risk profile.
- **NHANES I:** Features follow-up mortality data tracking participants from the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study.
- **Credit Card Fraud:** Contains transaction records categorized as either legitimate or fraudulent.
- **Bank Marketing:** Derived from a direct marketing initiative by a Portuguese banking institution, where the objective is to predict whether a client will open a term deposit account.

Experimental setup on rule-based models

To execute the rule-based evaluations, we adopt the experimental protocols established in [\[4\]](#). During the data preprocessing phase, each dataset is partitioned into three independent segments: 60% of the data is allocated to train the black-box model (X_{bb}, y_{bb}), 20% is assigned to train GLocalX(X_{le}, y_{le}), and the remaining 20% is reserved as unseen validation data (X_{vl}, y_{vl}) to verify model performance.

GLocalX requires a trained black-box model and a local explanation framework to synthesize global rules. For this purpose, a random forest ensemble [\[22\]](#) serves as the target black-box model, generating predicted labels for the 20% training partition assigned to GLocalX. We use the LORE algorithm to extract individual local explanations for each instance within this partition. A conceptual overview of the required components is provided in [Figure 2](#).

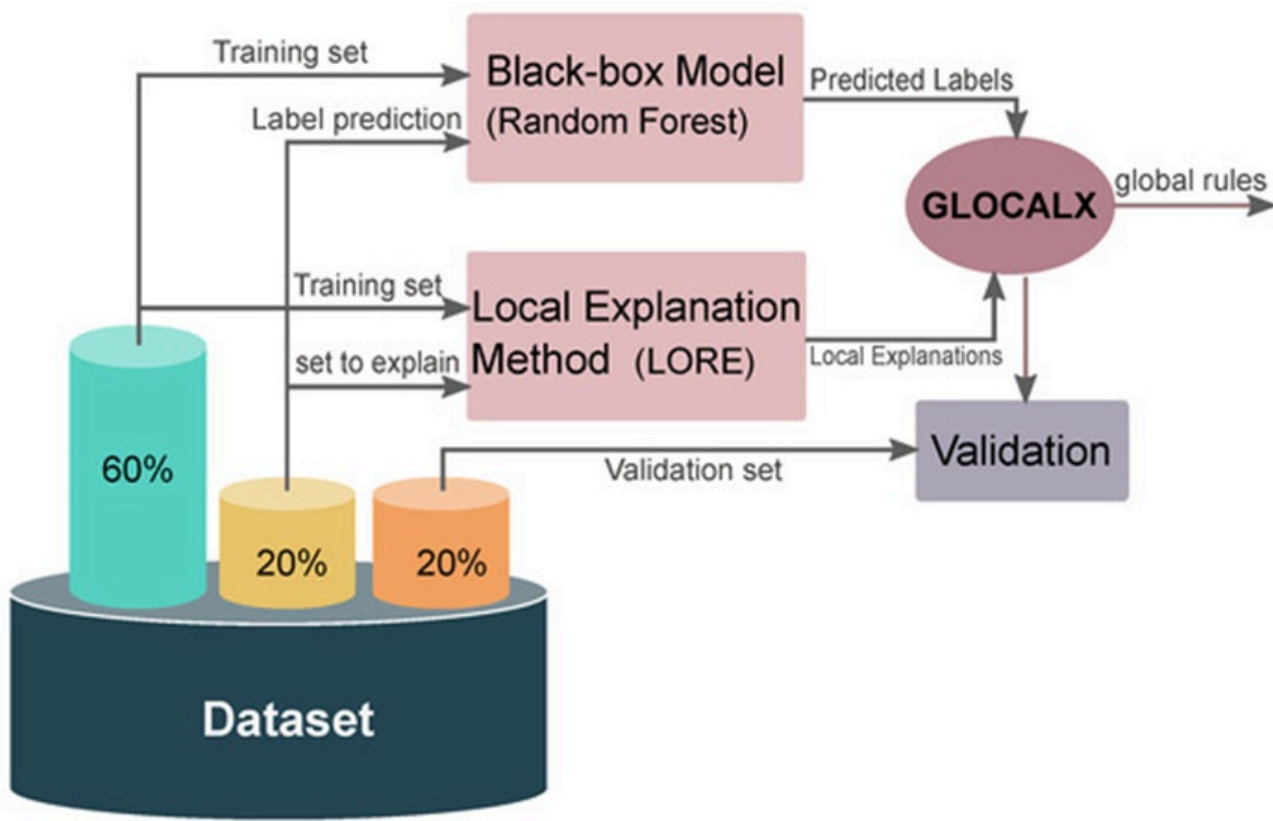


Figure 2. An overview of experiment setup for GLocalX.

For the evaluations involving JRip, the initial 80% of the dataset is used collectively as the training partition. It is worth noting that JRip internally splits its training data into a 2/3 growing set and a 1/3 pruning set, a mechanism that closely mirrors the internal partitioning utilized by GLocalX.

Because rule-based interpretability frameworks output discrete class predictions, model performance is assessed using accuracy and fidelity. Additionally, the total number of rules in the final model is tracked as a key metric for evaluating structural complexity.

Results and Discussion

This section presents the empirical findings obtained from the rule-based and feature-based modeling experiments.

Results on rule-based models

To assess the viability of using the transparent JRip framework as a substitute for the post-hoc explanatory model GLocalX, we evaluated GLocalX across a spectrum of values for its α parameter and compared the resulting rule sets with the rule set learned directly by JRip. **Table 1** outlines the performance profiles of JRip alongside the various α configurations of GLocalX on the adult dataset.

Table 1. Evaluation on adult dataset.

Algorithm	# Rules	Fidelity	Accuracy	α
GLocalX	96	0.856	0.723	95

	26	0.929	0.752	25
	10	0.942	0.762	10
	6	0.925	0.752	5
	2	0.911	0.729	2
JRip	28	0.957	0.839	

The empirical data indicates that on the adult dataset, GLocalX achieves its optimal trade-off between fidelity and rule base compactness at $\alpha=10$. Consequently, this specific configuration is selected for deeper analysis. Comparing the performance of GLocalX against JRip in Table 1 reveals that both frameworks achieve remarkably similar outcomes in terms of predictive accuracy and fidelity. The most precise rule base produced by GLocalX (at $\alpha=10$) yields a lower raw accuracy than JRip, though JRip produces a slightly more intricate rule structure.

Even when examining a GLocalX configuration with comparable structural complexity ($\alpha=25$), the core trend remains unchanged. Despite having no direct exposure to the underlying black-box model during training, JRip delivers an equally or more effective representation of its behavior. Specifically, it achieves a higher fidelity to the black-box system than GLocalX, despite the fact that GLocalX explicitly attempts to mimic the black-box outputs while JRip optimizes an entirely independent rule set.

Table 2 aggregates the experimental results across all evaluated datasets, utilizing the optimal α parameter custom-tuned for each respective data source. The consolidated metrics show that both algorithms maintain highly competitive performance across rule count and predictive accuracy, with JRip consistently retaining a minor advantage.

Table 2. Summary of Evaluation on datasets.

	JRip		GLocalX	
Dataset	Fidelity	Accuracy	Fidelity	Accuracy
<i>adult</i>	0.957	0.839	0.942	0.762
<i>German</i>	0.809	0.712	0.764	0.700
<i>NHANES I</i>	0.838	0.823	0.819	0.788
<i>compas-two-year</i>	0.836	0.782	0.804	0.756

To analyze the structural composition of these models, Table 3 contrasts the explicit rules extracted by GLocalX and JRip using the adult dataset. Although certain predictive features—such as "age", "capital-gain", and "marital-status"—appear consistently within the global rules of both GLocalX and JRip, the overall logical structures diverge substantially while maintaining nearly identical fidelity as global explanations.

This serves as a distinct manifestation of the Rashomon effect [20], which dictates that fundamentally different model configurations can capture identical predictive capacity, and that minor variations in a dataset can prompt significant structural shifts in learned symbolic models. Here, we observe this phenomenon from a novel perspective: completely distinct sets of logical rules can yield different explanations while providing the exact same level of fidelity to an underlying black-box architecture.

Table 3. Rules learned on adult.

(a) GLocalX
y>50k :- 41.0 <= age <= 49.0
hours_per_week >= 57.0
y>50k :- capital_gain <= 6808,
age >= 46.0.
.
.
.
y<=50k :- otherwise.
(b) JRip
y>50k :- marital-status = Married-civ-spouse
capital-gain >= 5178
y>50k :- marital-status = Married-civ-spouse
education = Bachelors
occupation = Exec-managerial
.
.
.
y>50k :- marital-status = Married-civ-spouse
age >= 29
hours-per-week >= 38
occupation = Sales
education = Masters
age <= 55
y <= 50k :- otherwise

Furthermore, the standard interpretability-accuracy trade-off implies that highly interpretable models often experience a degradation in raw predictive accuracy. Comparing the rule structures yielded by GLocalX and JRip regarding total rule length and the volume of antecedents per rule indicates that GLocalX generally constructs shorter rules with fewer constraints than JRip.

Consequently, as a direct inference, GLocalX may offer superior cognitive clarity for human interpreters by outputting more concise and less intricate rule sets. Conversely, JRip secures superior accuracy by mapping out more complex

logical pathways. It is important to emphasize that the interpretability and comprehensibility of rule bases can be analyzed from numerous conceptual dimensions and warrants deeper investigation.

Results on feature-based interpretability methods

As detailed in the preceding sections, GA²Ms and MAPLE were selected to evaluate feature-based interpretability. Here too, we aim to measure the performance of both techniques regarding their standalone predictive capacity (accuracy) alongside their alignment with the target black-box model (fidelity). We first examine the relative contribution of individual variables to the final output by plotting the feature importance hierarchies generated by both GA²Ms and MAPLE. Because the MAPLE algorithm does not natively output global feature importance as explicit weights, we utilize the average linear regression weights computed for each feature to establish its global importance ranking.

The complete metrics and graphical representations outlining the feature importance rankings for GA²Ms and MAPLE on the adult dataset are documented and illustrated in **Figure 3**. Comparing the outputs of MAPLE and GA²Ms confirms that variables such as "age" and "capitalgain" exert a substantial influence on the final predictions within both frameworks. However, specific features like "maritalstatus" display minimal importance inside MAPLE while acting as a primary driver within GA²Ms. This further illustrates the Rashomon effect, demonstrating that divergent feature weight allocations can be generated as valid explanations for the same underlying data distribution.

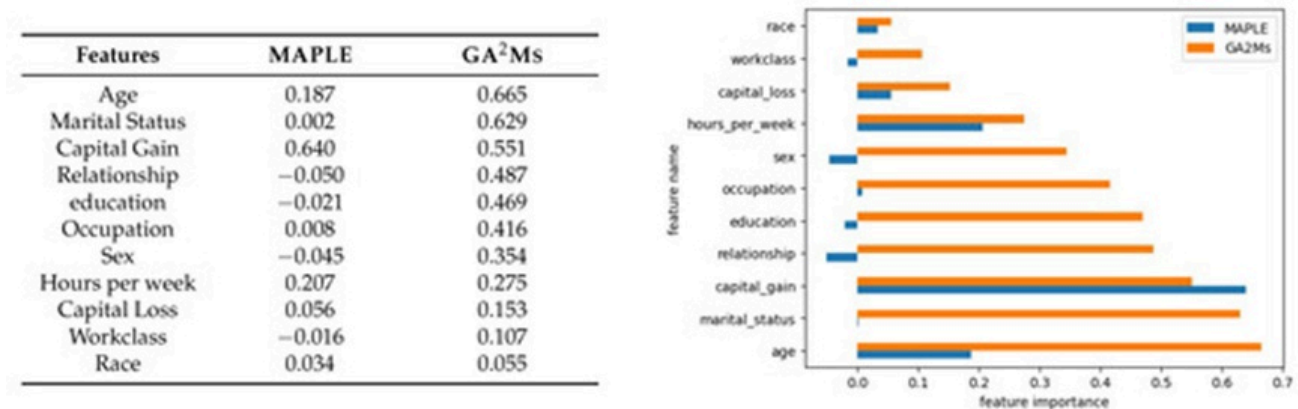


Figure 3. MAPLE and GA²Ms feature importance for adult dataset.

To ensure an equitable comparison of both techniques regarding fidelity, it is necessary to establish a uniform predictive mechanism for both MAPLE and GA²Ms. To accomplish this, we define the normalized feature importance weights extracted from GA²Ms and MAPLE as $\mathbf{W}_G \in \mathbb{R}^{m \times m}$ and $\mathbf{W}_M \in \mathbb{R}^{m \times m}$, respectively. For each explanation framework, we then apply **Eq. 4** to compute the corresponding value for every individual instance in the dataset:

$$\begin{aligned} v_{G,i} &= \mathbf{W}_G \cdot \mathbf{x}_i \\ v_{M,i} &= \mathbf{W}_M \cdot \mathbf{x}_i \end{aligned} \quad (4)$$

Within **Eq. 4**, the specific weights derived from MAPLE and GA²Ms are multiplied by the corresponding feature values for each sample in the dataset. To transform these continuous probability estimates into discrete binary labels, we optimize the decision threshold th using Receiver Operating Characteristic (ROC) curve analysis, as defined in **Eq. 5**:

$$th = \arg \min_p |TPR(p) + FPR(p) - 1|, \quad (5)$$

where TPR and FPR represent the true positive rate and false positive rate, respectively. The resulting values are subsequently utilized to compute the Area Under the Curve (AUC) to evaluate overall accuracy and fidelity.

Using this evaluation pipeline, the accuracy and fidelity metrics for both frameworks were calculated across the various datasets, with the final values summarized in [Table 4](#). The empirical results confirm that GA²Ms consistently outperforms MAPLE in both accuracy and fidelity across all evaluated datasets. This strongly reinforces the core hypothesis of this study: directly trained interpretable models can serve as highly effective explanations for black-box systems, even without direct access to them, because well-optimized transparent and black-box models will fundamentally correlate with one another due to learning from identical underlying data spaces.

Table 4. GA²Ms and MAPLE fidelity and accuracy evaluation on datasets.

	MAPLE		GA ² Ms	
Dataset	Fidelity	Accuracy	Fidelity	Accuracy
<i>Adult</i>	0.734	0.723	0.901	0.851
<i>German</i>	0.760	0.609	0.764	0.652
<i>Heart Disease</i>	0.801	0.734	0.813	0.789
<i>Credit card fraud</i>	0.955	0.947	0.989	0.939
<i>NHANES I</i>	0.788	0.745	0.867	0.826
<i>Bank</i>		0.713	0.807	0.731

Conclusion

The field of interpretable machine learning has become increasingly critical across a wide range of domain applications. The foundational principle underlying many black-box explanation methods is to approximate complex decision spaces globally or locally via an interpretable surrogate model. However, this approximation pipeline frequently sacrifices a significant portion of the original model's predictive capability, and determining whether the resulting surrogate is truly faithful to the primary black-box system remains a challenge. In this study, we demonstrated that, perhaps counterintuitively, transparent models trained entirely independently of the black-box system can achieve a level of fidelity to that black-box system that matches or exceeds the performance of post-hoc surrogates explicitly trained to mimic it.

Specifically, we evaluated GLocalX and JRip as representative post-hoc and inherently interpretable rule-based techniques, while utilizing MAPLE and GA²Ms to represent post-hoc and inherently interpretable feature-based methods. The experimental findings indicate that directly trained transparent frameworks deliver accuracy and fidelity metrics on par with advanced post-hoc explanation pipelines. However, the explicit explanations generated by these two paradigms—whether formatted as symbolic rule sets or continuous feature importances—can diverge significantly. Consequently, inherently interpretable models can be directly deployed as viable alternatives to post-hoc explanation methods. Furthermore, analyzing the structural variations between explanations yielded by rule-based versus feature-based methodologies presents an important direction for future research to optimize explanation efficiency and identify the most appropriate interpretability framework for a given dataset.

Acknowledgements

Supported by Johannes Kepler Open Access Publishing Fund and the federal state of Upper Austria.

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Aug 2024

Revised: 23 Sep 2024

Accepted: 06 Nov 2024

Published online: 15 March 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv.* 2018;51(5):93.
<https://doi.org/10.1145/3236009>.
- Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York (NY): ACM; 2016. p. 1135-44.
<https://doi.org/10.1145/2939672.2939778>.
- Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, editors. *Advances in Neural Information Processing Systems* 30. Red Hook (NY): Curran Associates; 2017. p. 4765-74.
- Setzu M, Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F. GLocalX: from local to global explanations of black box AI models. *Artif Intell.* 2021;294:103457.
<https://doi.org/10.1016/j.artint.2021.103457>.
- Lundberg SM, Erion GG, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2:56-67.
<https://doi.org/10.1038/s42256-019-0138-9>.

Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer; 2019. p. 193-209.
https://doi.org/10.1007/978-3-030-28954-6_10.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.
<https://doi.org/10.1038/s42256-019-0048-x>.

Molnar C. Interpretable machine learning [Internet]. 2019 [cited 2022 Sep 1]. Available from:
<https://christophm.github.io/interpretable-ml-book/>

Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. *Explainable AI: Interpreting, explaining and visualizing deep learning*. Cham: Springer; 2019. 11700 p.
<https://doi.org/10.1007/978-3-030-28954-6>.

Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, et al., editors. *Explainable and interpretable models in computer vision and machine learning*. Cham: Springer; 2018.

Fürnkranz J, Gamberger D, Lavrač N. *Foundations of rule learning*. Berlin: Springer; 2012.

Cohen WW. Fast effective rule induction. In: Prieditis A, Russell S, editors. *Machine Learning Proceedings 1995*. San Francisco (CA): Morgan Kaufmann; 1995. p. 115-23.
<https://doi.org/10.1016/B978-1-55860-377-6.50023-3>.

Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York (NY): ACM; 2013. p. 623-31.
<https://doi.org/10.1145/2487575.2487579>.

Guidotti R, Monreale A, Giannotti F, Pedreschi D, Ruggieri S, Turini F. Factual and counterfactual explanations for black box decision making. *IEEE Intell Syst*. 2019;34(6):14-23.
<https://doi.org/10.1109/MIS.2019.2957223>.

Plumb G, Molitor D, Talwalkar A. Model agnostic supervised local explanations. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. *Advances in Neural Information Processing Systems 31*. Red Hook (NY): Curran Associates; 2018. p. 2520-9.

Fürnkranz J. From local to global patterns: evaluation issues in rule learning algorithms. In: Morik K, Boulicaut JF, Siebes A, editors. *Local Pattern Detection*. Berlin: Springer; 2005. p. 20-38.
https://doi.org/10.1007/3-540-31814-3_2.

Knobbe AJ, Crémilleux B, Fürnkranz J, Scholz M. From local patterns to global models: the LeGo approach to data mining. In: Knobbe AJ, editor. *From Local Patterns to Global Models: Proceedings of the ECML/PKDD-08 Workshop (LeGo-08)*. Darmstadt: University and State Library Darmstadt; 2008. p. 1-16.

Yang C, Rangarajan A, Ranka S. Global model interpretation via recursive partitioning. In: *Proceedings of the 20th IEEE International Conference on High Performance Computing and Communications; the 16th IEEE International Conference on Smart City; the 4th IEEE International Conference on Data Science and Systems*. Piscataway (NJ): IEEE; 2018. p. 1563-70.

van der Linden I, Haned H, Kanoulas E. Global aggregations of local explanations for black box models. *arXiv [Preprint]*. 2019. arXiv:1907.03039.

Breiman L. Statistical modeling: the two cultures. *Stat Sci*. 2001;16(3):199-231.

Dua D, Graff C. UCI machine learning repository. Irvine (CA): University of California, Irvine, School of Information and Computer Sciences; 2017 [cited 2023 Apr 3]. Available from: <https://archive.ics.uci.edu/ml>

Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
<https://doi.org/10.1023/A:1010933404324>.