All Models Are Wrong, but Some Are Interchangeably Right

Anonymous Authors[1]
[1]Anonymous Affiliations
{authors}@anonymous.com

Abstract

Many tasks require practitioners to select machine learning (ML) models among many possible models that differ in terms of inductive biases, computational costs, and interpretability. At the same time, real-world datasets often contain underlying properties that are difficult to uncover but can significantly influence the predictive performances of the models. Overall, this raises an important question: **to what extent do different models agree in what they reveal about the underlying phenomena embedded in the dataset at hand?** In this work, we explore this question by analyzing the agreement between ML models using attributive post-hoc explainability methods. Specifically, we leverage Shapley-based feature-importance rankings to measure the similarity of explanations across diverse model families, both at a local (instance level) scale and at the global (feature level) scale. We interpret these post-hoc attributions as proxies for what ML models have learned from the data, and for the extent to which different models capture and agree on the same underlying phenomena. Our study covers 30 benchmark problems in regression and classification to provide a systematic comparison of 46 predictive ML models. We identify recurring agreement regimes in which models with very different architectures produce highly similar top-ranked explanations. We further identify model families that consistently exhibit such agreements. Additionally, our analysis uncovers systematic interactions between certain model families and the datasets or tasks they are applied to. Low-agreement regimes are associated with unstable explanations, meaning that both task and dataset characteristics are important factors to consider when selecting models. Finally, we find that some lightweight models can generate explanations that are effectively interchangeable with those of more complex architectures, while requiring fewer computational resources or less data. In this context, explanation agreement provides a practical criterion that supports automated XAI (AutoXAI) and model recommendation workflows.

1 Introduction

ML models are increasingly deployed in high-stakes decision-making pipelines, including healthcare, criminal justice, and credit lending, where understanding the rationale behind predictions is critical [Rudin, 2019]. In such contexts, it is no longer sufficient for a model to merely maximize predictive accuracy; it must also provide an auditable account of its reasoning [Lakkaraju and Rudin, 2017]. Yet, many state-of-the-art system such as deep neural networks and ensemble methods, remain opaque, raising concerns about trust, accountability, and regulatory compliance [Apley and Zhu, 2020]. Explainable Artificial Intelligence (XAI) has emerged to address this tension. Among existing techniques, SHapley Additive exPlanations (SHAP) have become a widely adopted framework for post-hoc feature attribution, offering a principled approach to interpreting complex models [Lundberg and Lee, 2017]. Implicit in much of this practice is the assumption that different model architectures yield meaningfully different explanations, making interpretability inherently model-dependent. However, recent work suggests that the presence of an explanation does not guarantee its reliability [Ribeiro *et al.*, 2016]. The field has shifted from a narrow focus on explanatory faithfulness toward the broader question of **explanatory consistency** [Rudin, 2019]. Central to this issue is the **Rashomon Effect**: multiple models can achieve similar predictive performance while relying on different internal decision mechanisms [Breiman, 2001]. This multiplicity can lead to contradictory explanations for the same task, a challenge often framed as the **disagreement problem** [Krishna *et al.*, 2022]. While prior studies have documented cases of disagreement, systematic empirical evidence across a broad spectrum of models remains limited.

In this work, we adopt a complementary perspective: rather than asking when explanations diverge, we ask **when they converge, and what such convergence implies about the relationship between data, models, and interpretability**. Specifically, we analyze the alignment of SHAP-based feature importance rankings across a diverse set of models. If explanations vary widely with model choice, interpretability must be treated as model-contingent. Conversely, if explanations remain stable across architectures, this suggests a form of descriptive robustness rather than model-specific variability.

To structure this analysis, we distinguish between **local** and **global** explanations. Local explanations assign feature importance to individual predictions, enabling instance-level interpretability, whereas global explanations aggregate attribution patterns across an entire dataset, capturing more stable, population-level insights [Minh *et al.*, 2022]. We also contrast **post-hoc** explanations from black-box models with **intrinsically interpretable** models, allowing us to assess whether observed patterns are driven by model structure or by the data itself [Bibal, 2020].

Experimentally, we conduct a large-scale study of SHAP-based feature importance rankings across and quantify cross-model alignment using Normalized Discounted Cumulative Gain (NDCG), a standard metric for ranking similarity [Järvelin and Kekäläinen, 2002]. Our results reveal a new pattern: despite substantial architectural differences, many models produce highly similar explanations. We identify clusters of models with strongly aligned feature rankings, indicating that in many cases, explanatory behavior is governed more by data structure than by model complexity. **To our knowledge, this is the first systematic benchmark that evaluates explanation agreement across a broad diversity of model families and tasks under a unified ranking-based metric.** Based on these findings, we introduce the notion of **explanation interchangeability**. Two models are deemed interchangeable when their SHAP rankings exhibit high consensus (around 90%) in instance-based feature rankings, indicating that the identity of the most influential features is largely insensitive to model choice under the considered explainer. This has important practical consequences: in many settings, lightweight models yield explanations that are statistically indistinguishable from those of far more complex ensembles while being dramatically cheaper to train and analyze. Consequently, heavy models often provide little additional interpretive value. This enables more principled model selection for explanation-focused workflows, including the identification of models that maximize explanatory consensus while minimizing computational cost. This is particularly relevant in regulated or resource-constrained settings, where organizations must justify model behavior while minimizing computational and operational costs

The contributions of this paper are threefold:

1. We present a large-scale systematic comparison of SHAP-based feature importance rankings across 46 predictive models and 30 real-world datasets.
2. We highlight recurring regimes of **explanation interchangeability**, demonstrating that interpretability is frequently driven more by data properties than by model architecture.
3. We derive practical guidelines for selecting computationally efficient surrogate models that preserve explanation reliability.

The remainder of this paper is organized as follows. Section 2 reviews related work on XAI, explanation disagreement, and the Rashomon effect. Section 3 describes the classification and regression benchmark datasets. Section 4 details our experimental protocol and cross-model NDCG analysis. Finally, Section 5 discusses the implications for robust interpretability and outlines future research directions.

2 Motivations and background

This section positions our work on post-hoc explainability and presents well-known sources of explanatory divergences, such as the **disagreement problem** or the **Rashomon effect**.

2.1 Current Trends and Problematics in Black-Box AI

A central tension in contemporary ML concerns the relationship between model expressivity, data availability, and interpretability [Garouani *et al.*, 2024]. In low-data or noise-dominated regimes, simple and inherently interpretable models, such as linear models or shallow decision trees, often provide a direct and intelligible mapping between inputs and outputs. Their explicit structure supports human understanding and qualitative reasoning, but this transparency is frequently obtained at the cost of limited representational power, reduced predictive accuracy, or instability under small data perturbations. In contrast, highly expressive models with many degrees of freedom, including multilayer perceptrons, deep neural networks, and large ensemble methods, are specifically designed to exploit large-scale datasets. These models can approximate complex, nonlinear decision boundaries and interactions, thereby achieving superior predictive performance, but they do so by encoding decision logic in high-dimensional parameter spaces that are largely opaque to human interpretation [Palar *et al.*, 2025]. This opposition reflects the epistemological divide described by [Breiman, 2001] between the **data modeling** and **algorithmic modeling** cultures. Data modeling relies on explicit assumptions about the stochastic data-generating process, producing simple, interpretable models that favor inference and explanatory insight over predictive accuracy. These models perform well with limited data or strong prior knowledge but often struggle with complex, high-dimensional phenomena. In contrast, algorithmic modeling treats the mechanism as unknown, focusing on flexible, data-driven prediction. While powerful for large-scale tasks, interpretability becomes a post-hoc concern rather than an inherent property of the model, giving rise to **post-hoc XAI**.

Importantly, this tension should not be interpreted as a universal or immutable trade-off between interpretability and accuracy. Rather, it is strongly conditioned on the data regime, the dimensionality of the feature space, and the inductive biases of the learning algorithm [Beckh *et al.*, 2021]. Recent empirical evidence has challenged the widespread assumption that black-box complexity is a prerequisite for high performance in tabular and structured data settings. Large-scale benchmarking studies [Christodoulou *et al.*, 2019] demonstrate that transparent algorithms, such as logistic regression, scoring systems, and rule-based models, often achieve predictive performance within approximately 5% of state-of-the-art black-box methods across a wide range of medical and tabular datasets [Peterson *et al.*, 2024]. These findings suggest that, in many practical scenarios, the use of highly opaque models is a default choice rather than a mathematical necessity, and resonate with benchmarking suites that show substantial variance across datasets and tasks [Olson *et al.*, 2017]. Notwithstanding, these results should be mitigated in more extreme regimes characterized by massive datasets, high-dimensional feature spaces, and complex hierarchical structures. In such cases, deep learning and flexible model architectures are generally more fitted, while transparent models often fail to scale or to capture the relevant interactions because their simplicity makes them too rigid or sensitive to the curse of dimensionality. Still, in these cases, a shift in paradigm has emerged: rather than replacing black-box models, ML models are leveraged to extract intelligible, faithful, and stable explanations from underlying models whose complexity appears unavoidable [Moss *et al.*, 2022]. In that case, ML models are treated as **surrogate models** [Saves *et al.*, 2024] and have fueled the rapid development of fast-to-evaluate post-hoc explainability methods, with SHAP, LIME, or gradient-based approaches becoming de facto standards in applied ML [Sundararajan *et al.*, 2017].

Despite their theoretical appeal and growing adoption, post-hoc explanation methods have exposed a fundamental fragility in current XAI practice, commonly referred to as the **Disagreement Problem** [Krishna *et al.*, 2022]. Empirical studies reveal that different explanation techniques applied to the same trained model often produce conflicting feature rankings, and that even minor changes in model initialization, training data splits, or random seeds can result in substantially different explanations for models with indistinguishable predictive performance. This instability undermines the use of explanations as reliable scientific or decision-support tools and motivates rigorous analyses of explanation variability [Müller *et al.*, 2023]. Two primary sources of explanation variability can be distinguished. The first is method-driven variability, which arises from the explanation algorithm itself. Many post-hoc methods rely on stochastic sampling, surrogate fitting, or heuristic design choices, making them sensitive to hyperparameters such as kernel widths, baseline selection, sampling budgets, or neighborhood definitions; smoothing and aggregation strategies (e.g., SmoothGrad) mitigate but do not eliminate such sensitivity [Smilkov *et al.*, 2017]. The second, more fundamental source is model-driven variability, which is rooted in predictive multiplicity and formalized by the **Rashomon Effect**. As articulated in [Anderson, 2016] and empirically analyzed in [Müller *et al.*, 2023], the hypothesis space often contains a large set of models, referred to as the Rashomon set, that achieve near-identical empirical risk while relying on different combinations of features, interactions, or internal representations. When feature redundancy or multicollinearity is present, attributions can be arbitrarily partitioned among correlated predictors. This phenomenon has profound implications for explainability. Two models drawn from the Rashomon set may be fully interchangeable from a predictive standpoint, yet yield explanations that are descriptively incompatible. In such cases, explanations reflect the arbitrary outcome of the optimization process rather than a unique, data-driven relationship between inputs and outputs. In high-stakes contexts, this instability poses a critical risk: if retraining a model with a different random seed shifts the dominant explanatory factor from one variable to another, the resulting explanation cannot be interpreted as a robust insight, let alone a causal mechanism [Mehdiyev *et al.*, 2025]. Instead, it becomes an artifact of model selection within a vast space of equally performant alternatives. Similarly, recent work on Shapley effects and dependence-aware attribution methods shows that, in the presence of correlated inputs, feature attributions are not uniquely defined because the underlying cooperative game depends on how feature dependence is modeled. While Shapley-based decompositions (including interaction or dividend-based formulations) provide a principled allocation of shared effects once a value function is specified, the choice of marginalization or conditioning, implicitly fixing a joint input distribution, ultimately determines how importance is assigned among correlated predictors; without such assumptions or an explicit causal model, attribution remains fundamentally underdetermined [Idrissi *et al.*, 2021].

Beyond variability, several additional problems constrain the operational utility of black-box explanations. First, computational and statistical costs are non-trivial: exact Shapley computations are combinatorial, and practical algorithms trade off bias, variance, and computational tractability, which harden large-scale deployment [Lundberg and Lee, 2017]. Second, many attribution methods implicitly rely on conditional expectation estimators; when these estimators are misspecified or when covariates are strongly dependent, attributions can be misleading unless dependence is explicitly modeled [Idrissi *et al.*, 2021]. Third, explanation fragility to small input perturbations or adversarial manipulations raises concerns about faithfulness and safety in operational systems [Smilkov *et al.*, 2017]. Finally, the evaluation of explanations is itself an open problem: while rank-based measures (Kendall's τ, Spearman ρ, NDCG) and fidelity metrics are useful for cross-model comparison, no single metric captures human comprehensibility, causal validity, or task-specific utility simultaneously [Wang *et al.*, 2023]. The literature has responded with several converging strategies. Advocates of inherently interpretable models argue for preferring transparent models by design in high-stakes settings and have produced practical high-fidelity interpretable architectures and scoring systems [Caruana *et al.*, 2015]. Complementary lines of work pursue **regularization for explainability**, whereby training-time penalties encourage sparse, stable, or human-aligned representations so that post-hoc attributions become more consistent and meaningful [Plumb *et al.*, 2020]. Ensemble and consensus techniques quantify explanation uncertainty by aggregating attributions across model ensembles or across explainers and reporting agreement intervals or rank-consensus statistics [Levy *et al.*, 2025]. Surrogate-model workflows construct compact, interpretable proxies to summarize global behavior while reserving local post-hoc tools for detailed inspection. Dependence-aware attribution and causal conditioning operationalize this trade-off by explicitly encoding the priors or structural constraints that resolve attribution ambiguity: they formalize choices about marginalization, conditioning, or causal structure that effectively reduce the space of admissible explanations [Kennedy and O'Hagan, 2001]. In that sense, these methods are between the two strategies described above, aligning with interpretable-by-design and regularization-for-explainability approaches (which impose structural or sparsity priors during training) while remaining compatible with surrogate and ensemble workflows [Rudin, 2019]. Practically, such priors should be elicited from domain knowledge and made explicit so their influence on attributions can be assessed and validated within a Bayesian or uncertainty quantification framework [Livet and Varenne, 2020; Wilhelm and Zweig, 2024]. Together with ensemble/consensus techniques that quantify explanation uncertainty and surrogate proxies that capture global behavior, dependence-aware attributions form a coherent toolbox for producing explanations that are both principled and practically actionable [Lakkaraju and Rudin, 2017]. Taken together, these trends indicate that black-box XAI is maturing: the field is shifting from single-method visualizations toward characterizing explanation distributions and their dependence on model choice, data regime, and estimation procedure [Apley and Zhu, 2020]. In practice, this encourages preferring interpretable-by-design models when stakes are high, evaluating explanations with multiple metrics and human-grounded protocols, accounting for dependence and causal structure, and reporting uncertainty or consensus rather than a single attribution vector [Löfström *et al.*, 2022].

Our study is situated within this evolving landscape. Motivated by the Rashomon-driven instability of attributions and the operational need for robust interpretability, we adopt an interchangeability perspective: we quantify consensus across model families using rank-based and top-weighted metrics and investigate the conditions under which explanations reflect genuine data signal versus model assumptions. This approach extends recent benchmarking and consensus work [Le *et al.*, 2023] and contributes operational criteria for trustworthy interpretation in regimes where predictive multiplicity is unavoidable. From an evaluation perspective, it is crucial to distinguish between **explanation fidelity** and **human interpretability**, two concepts often conflated in XAI [Doshi-Velez and Kim, 2017]. Fidelity measures how accurately an explanation reflects the behavior of the underlying predictive model, while interpretability captures the extent to which the explanation can be meaningfully understood and acted upon by a human. Post-hoc methods such as SHAP primarily optimize fidelity, producing local attributions that, in expectation, align with the model's input–output mapping [Lundberg and Lee, 2017]. Yet, high-fidelity explanations are not necessarily stable, sparse, or cognitively tractable, especially under feature dependence or predictive multiplicity. Conversely, explanations designed for human interpretability, such as simplified rule lists or sparse scoring systems [Mehdiyev *et al.*, 2025], may trade off fidelity to individual black-box predictions. This tension motivates a shift from single-model explanations toward assessing robustness and consensus across families of near-optimal models, evaluating fidelity collectively rather than on a per-instance basis.

2.2 Mathematical Framework for Post-hoc Explainability

To systematically study how different ML models yield explanations, we formalize post-hoc interpretability in terms of feature attribution vectors. Let $\hat{f} : \mathcal{X} \to \mathbb{R}$ denote a predictive model trained on d features, and let $\mathbf{x} \in \mathcal{X}$ be an input instance. A post-hoc explainer assigns to each feature i a contribution $\phi_i(\mathbf{x}; \hat{f})$, capturing its influence on the model's output. The resulting attribution vector for $\mathbf{x}$ is $\mathbf{s}(\mathbf{x}; \hat{f}) = (\phi_1, \ldots, \phi_d) \in \mathbb{R}^d$, which provides a compact representation of local explanations suitable for quantitative comparison across models.

Among post-hoc methods, SHAP stands out due to its theoretical foundations rooted in cooperative game theory. For a subset of features $S \subseteq \{1, \ldots, d\}$, the SHAP value function is defined as $v_{\mathbf{x}}(S) = \mathbb{E}[\hat{f}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$, which represents the expected output of $\hat{f}$ when features in S are fixed to their observed values, while others vary according to the underlying distribution. The SHAP attribution for feature i is computed as the average marginal contribution across all feature subsets: $\phi_i(\mathbf{x}) = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{i\}} \frac{|S|!\,(d-|S|-1)!}{d!} \left(v_{\mathbf{x}}(S \cup \{i\}) - v_{\mathbf{x}}(S) \right)$, ensuring the additive decomposition: $\hat{f}(\mathbf{x}) = v_{\mathbf{x}}(\varnothing) + \sum_{i=1}^{d} \phi_i(\mathbf{x})$. Exact computation is combinatorial and often intractable; practical algorithms such as TreeSHAP, KernelSHAP, or Monte Carlo approximations introduce computational and statistical biases, and rely on explicit assumptions about background distributions or feature dependence [Apley and Zhu, 2020]. Variants such as dependence-aware Shapley and Shapley effects further formalize the role of correlations and conditional distributions, highlighting that explanations are only fully defined once these assumptions are specified.

Because feature attributions vary in scale across model architectures (e.g., logits in neural networks versus probability masses in tree ensembles), comparing raw attribution magnitudes is often misleading. Rank-based metrics therefore provide a more robust notion of agreement by emphasizing the consistency of top-ranked features.

Let $|S_i^{(k)}| \in \mathbb{R}^d$ denote the vector of absolute SHAP values for instance k under model M_i, and let $\pi_i^{(k)}$ be the ranking of features induced by sorting $|S_i^{(k)}|$ in decreasing order. Given two models M_i and M_j, the DCG of the ranking $\pi_i^{(k)}$ evaluated using relevance scores from M_j is defined as $\mathrm{DCG}(\pi_i^{(k)}, M_j) = \sum_{r=1}^{d} \frac{\mathrm{rel}^{(k)}_{\pi_i^{(k)}(r)}}{\log_2(r+1)}$, where $\mathrm{rel}^{(k)}_{\ell} = |S_j^{(k)}[\ell]|$ denotes the relevance of feature ℓ according to model M_j. The corresponding normalized score is obtained by dividing by the ideal DCG, i.e., the DCG obtained when ranking features according to M_j itself: $\mathrm{NDCG}(|S_i^{(k)}|, |S_j^{(k)}|) = \frac{\mathrm{DCG}\left(\pi_i^{(k)}, M_j\right)}{\mathrm{DCG}\left(\pi_j^{(k)}, M_j\right)}$. To aggregate instance-level comparisons, we define the instance-averaged directional similarity between models M_i and M_j as $\rho_{ij}^{\mathrm{NDCG}} = \frac{1}{n} \sum_{k=1}^{n} \left(\mathrm{NDCG}(|S_i^{(k)}|, |S_j^{(k)}|) \right)$, where n is the number of test instances. Values of $\rho_{ij}^{\mathrm{NDCG}} \approx 1$ indicate strong agreement in top-feature rankings, suggesting that the dominant explanatory patterns are driven by the underlying data rather than by model-specific properties. Note that $\rho_{ij}^{\mathrm{NDCG}}$ is directional and generally differs from $\rho_{ji}^{\mathrm{NDCG}}$. This formalism allows us to operationalize the notion of **explanation interchangeability**. By representing models through their attribution vectors $\mathbf{s}(\mathbf{x}; \hat{f})$ and comparing them using $\rho_{ij}^{\mathrm{NDCG}}$, we can identify clusters of models within the Rashomon set whose explanations are consistent despite architectural differences. Such clusters correspond to scenarios where the top-ranked features are largely dictated by the structure of the data, revealing that interpretability often reflects stable, intrinsic properties of the data-generating process rather than the choice of predictive model.

Moreover, this framework enables several practical insights. First, it provides a principled metric for selecting a **centroid model** that maximizes explanatory consensus within a family of high-performing predictors, reducing computational cost without compromising interpretability. Second, it quantifies conditions under which model interchangeability fails, for instance, in datasets with strong nonlinear interactions, high feature redundancy, or when models have divergent inductive biases. Finally, by formalizing the link between data structure, predictive multiplicity, and feature attribution agreement, this approach grounds our subsequent experimental analysis, allowing us to systematically explore the interplay between accuracy, explanation stability, and computational efficiency across diverse ML architectures.

3 Datasets and ML models

To evaluate our approach, we selected 30 datasets from the PMLB benchmark [Olson *et al.*, 2017], and removed duplicate inputs. The ML models are given in Figure 1. We separate the datasets according to their targets into regression and classification tasks. These datasets cover a broad range of applications, as well as combinations of categorical, ordinal, and continuous features. The descriptions of the selected datasets and models are given in the supplementary materials. For each dataset, we use a random 70%/30% train/test split to evaluate predictive performance and explanation agreement. Every ML model comes with a k-folds cross validation of the hyperparameter optimization to ensure a fair comparison between the best versions of the ML for every dataset.

3.1 Classification

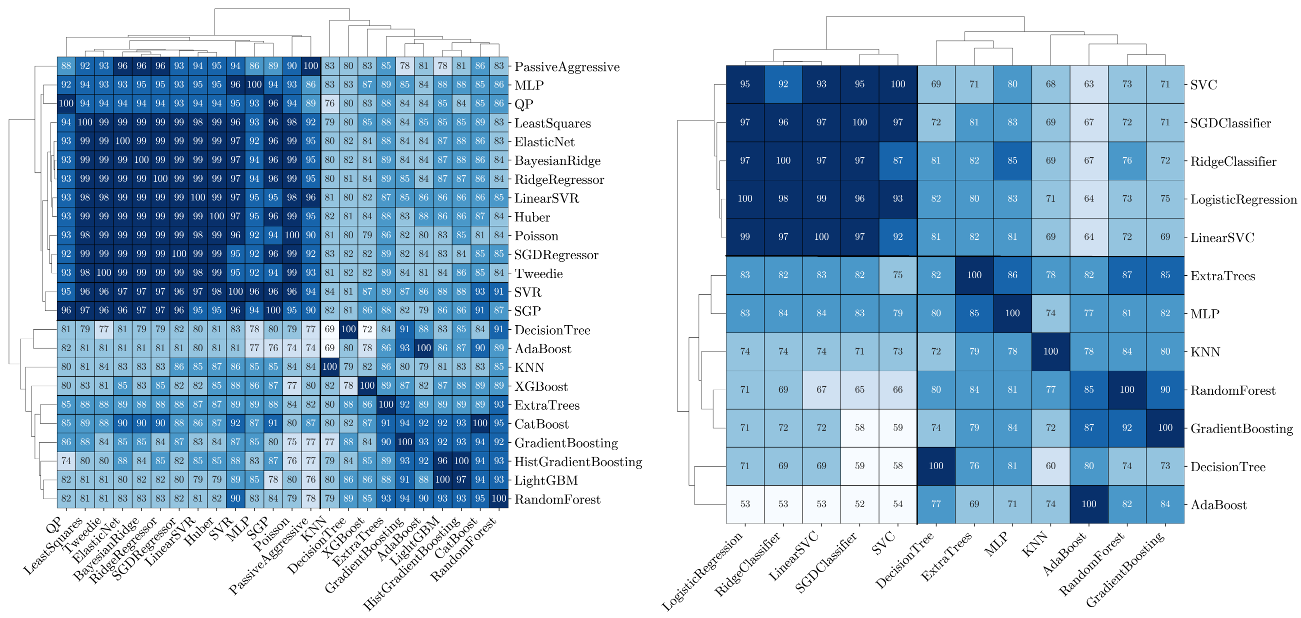
We evaluate 15 classification ML models (detailed in the supplementary materials) on 15 classification datasets, covering both binary and multi-class problems. These datasets span a wide range of sizes, dimensionalities, and numbers of classes. The number of instances ranges from 150 to nearly 50,000, with feature dimensions varying from 5 to 61 and up to four target classes. Handling explanations in multi-class settings is inherently more complex than in binary classification or regression because SHAP returns one attribution vector per class for each instance. To maintain a consistent and tractable experimental protocol across datasets and models, we adopt a pragmatic strategy: for each instance, we retain only the SHAP attributions associated with the class predicted by the model. This enables the construction of a single explanation matrix per dataset and simplifies the computation of global agreement metrics between models. However, this simplification increases the dependence of explanations on model behavior and predictive accuracy rather than solely on the underlying data distribution: misclassified instances are explained with respect to an incorrect class, and classes that are harder to predict tend to be underrepresented in global visualizations. To evaluate the robustness of our conclusions, we experiment with 16 machine-learning models on the classification datasets and extend the analysis to the regression cases, where explanations are single-output, and comparisons are less prone to the bias introduced by predicted-class selection. For classification, GaussianNB, LightGBM, CatBoost, and XGBoost proved prohibitively slow on several many-class datasets. Furthermore, we removed four datasets from the global analysis: Adult and Shuttle due to their large size (over 48,000 instances), and cmc and wine, which caused failures or instability for multiple models.

3.2 Regression

We evaluate 31 ML regression models over 15 Regression test cases. We categorize these regression datasets by size according to the number of samples that they contain: **small** (0 to 200 points), **medium** (500 to 4000 points), **big** (4000 to 10000), and **very big** (10000 to 40000). For medium datasets, the RMTS and TabPFN models are too costly (more than 10^5 seconds) to be tested for more than a few hundred data points or more than 5 variables. For big and very big datasets, we cannot reasonably test the most computationally expensive model. Therefore, we removed IDW, RBF, PCE, CIEL, and KPLS. For very large datasets, we do not allow SVM to use a RBF kernel, as the computing burden would be too great. All the models and datasets operations are similarly seeded to allow for the replication of the results. To quantify the quality of the explanations, we compute NDCG over each model's SHAP-value importance ranking in both the regression and classification tasks [Burges *et al.*, 2005]. Note that these importance values are limited to their absolute influence and that other metrics taking into account the sign of the influence, such as the Composition of Rank, Influence, and Accuracy described in [Wang *et al.*, 2023], may lead to different results.

4 Experimental results

To quantify the quality and consistency of model explanations, we compute the NDCG over each model's SHAP-based feature-importance ranking for both regression and classification tasks [Burges *et al.*, 2005]. In practice, we rely on the automatic option of the `shap` Python library [Lundberg and Lee, 2019], which adaptively selects among seven approximation techniques instead of computing exact SHAP values. We use NDCG instead of classical rank correlations because it emphasizes agreement among top-ranked features, reflecting practical interpretability where only a few features matter; similar trends are observed with Spearman. We emphasize that these importance scores rely only on the absolute magnitude of SHAP values; consequently, alternative metrics that also account for the direction of influence, such as the Composition of Rank, Influence, and Accuracy proposed in [Wang *et al.*, 2023], could lead to different conclusions. As discussed in Section 2.2, the resulting NDCG matrices are non-symmetric because the normalization depends on the reference model used to compute the ideal ranking. Nevertheless, large discrepancies between $\rho_{ij}^{\mathrm{NDCG}}$ and $\rho_{ji}^{\mathrm{NDCG}}$ indicate that two models rank features very differently, provid-



(a) Median agreements for Regression.

(b) Median agreement for Classification.

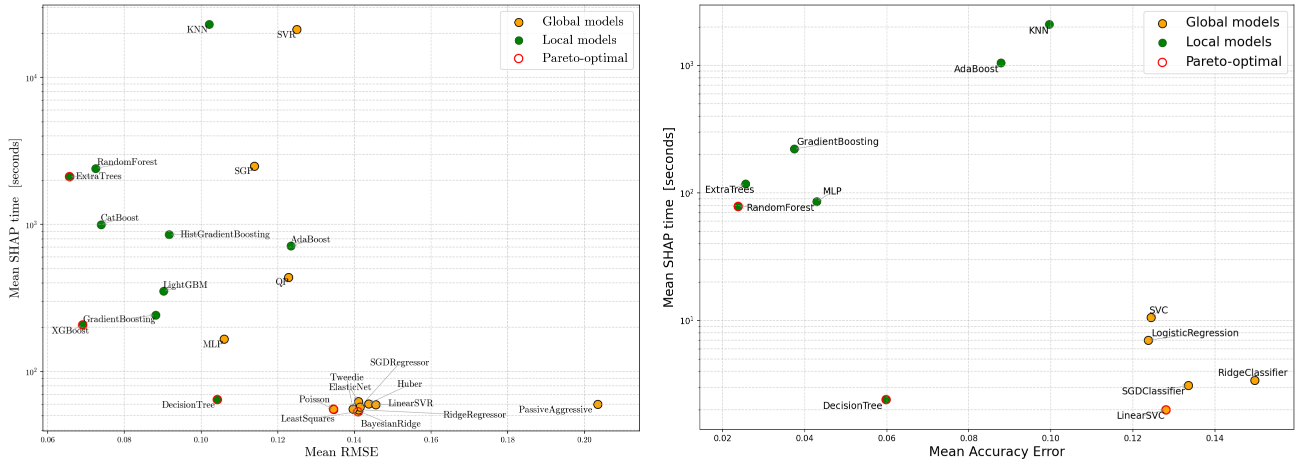
Figure 1: Median NDCG agreements for the 15 regression and 11 classification datasets.

ing an additional implicit measure of model dissimilarity captured in our analyses. To ensure reproducibility, every model and data split is initialized with the same random seed. Figure 1 reports the pairwise median agreements computed on the benchmark problems' validation sets. Supplementary materials and code are open-source and open-data on our online repository[1].

The SHAP-based NDCG analysis reveals two main subgroups of models. Each subgroup exhibits strong internal agreement and weaker inter-group similarity, suggesting two distinct families of ML models with different explanation patterns: broadly speaking, more **local** versus more **global** models. The first group, shown in the upper left, contains smooth models that perform well in familiar settings but struggle to generalize to broader, unseen patterns due to strong structural or parametric rigidities. Their SHAP explanations primarily reflect data tendencies and structural regularizations. For instance, this "global" group includes kernel-based models along with linear and quadratic models. In contrast, the second group, shown in the lower right, includes models that better capture complex relationships, but may also fit noisy data rather than the underlying mechanisms. Tree-based ensembles tend to be locally sensitive when their structural complexity is high, while neighborhood-based methods are locally sensitive to the sampled data. These results illustrate a practical manifestation of the **no free lunch theorem**: no single model is optimal across all datasets, and different inductive biases lead to systematically different explanation patterns [Lattimore and Hutter, 2013]. A finer-grain analysis is provided in the online Supplementary Material, where more models are considered, and the analysis is repeated for different dataset sizes. Notably, the first separation in the dendrogram — between the two largest blocks — explains roughly 30–60% of the total variance between the explanations of our 46 models, as shown by the dendrograms adjacent to the NDCG similarity matrices [Cormack, 1971]. We emphasize that these are median results and that the level of agreement can vary across datasets. In the Supplementary Material, we further apply our methodology to a real-world problem, namely Schelling's segregation model [Schelling, 1969], for both regression and classification tasks. In this setting, we observe stronger inter-group correlations and a clearer separation between model families. Taken together, these findings suggest that explanation choice and ML model selection should be tailored to the specific dataset and task, rather than relying on any single model, highlighting the importance of systematically evaluating multiple surrogates to obtain robust, informative explanations.

Following the discovery of these two groups of highly similar machine-learning models in terms of post-hoc SHAP explanations, Fig. 2 shows computational time versus RMSE for the models that ran in less than 10^5 seconds (time is plotted on a log scale for readability). Overall, the locally focused models such as ExtraTrees or Random Forest appear Pareto-optimal when high local precision is required, which is consistent with their propensity to overfit. Conversely, the Generalized Linear Model (Poisson) and the simple ordinary least-squares linear regression or Linear Support Vector Machine are Pareto-optimal choices when global trends must be captured quickly. Within the local group, models with lower median agreement also exhibit poorer accuracy and higher RMSE. Agreement is higher on simpler datasets than on more complex ones, suggesting a positive association between explanation agreement and predictive performances, even if the

[1] Anonymous: https://github.com/sub5716ijcai2026/RecoSHAP



(a) Computational time versus RMSE for Regression datasets. (b) Computational time versus Accuracy for Classification datasets.

Figure 2: Computational time versus errors for the 15 regression and 11 classification datasets.

DecisionTree is a good trade-off thanks to its ability to deliver fast explanations. These findings offer practical guidance for selecting surrogate models in ML for post-hoc XAI, depending on whether regression precision, classification reliability, uncertainty calibration, or interpretable explanations are the primary objectives. For instance, in our experiments, the nonlinear Support Vector Machine (SVM) explanations were 400 times slower to compute while exhibiting a similarity of about 95% compared to linear regression, yet they reduced RMSE by only about 14%. Hence, SVMs may be a suboptimal choice when explanation cost is a concern. These results can also be interpreted as a limitation of SHAP and an argument for alternative or complementary explanation methods (e.g., LIME [Ribeiro *et al.*, 2016]) or for interpretable-by-design (ante-hoc) models, since SHAP often reflects ML models' behavior more than the underlying data-generating mechanism [Retzlaff *et al.*, 2024].

5 Conclusions and Perspectives

We studied the extent to which post-hoc SHAP explanations agree across a wide range of machine-learning models and datasets using NDCG as a rank-based similarity measure. Our analysis reveals that explanations exhibit clear clustering patterns, broadly separating models with global, smooth inductive biases from those driven by local, data-dependent behavior. The main dendrogram split explains a substantial fraction of explanation variance, supporting the practical notion of **explanation interchangeability** within model families. We also observe a clear trade-off between predictive accuracy and explanation cost: tree ensembles typically achieve lower RMSE but require much more expensive SHAP computations, whereas simple parametric models are fast to explain but can underfit complex patterns. Crucially, we find that SHAP explanations often reflect the ML model as much as the underlying data-generating process. When multiple near-optimal models exist (Rashomon sets), explanations can be unstable across model classes, initializations, or data splits, cautioning against treating any single attribution as definitive evidence of causality. Overall, explanation agreement depends on both the model and the data: patterns are visible across tabular benchmarks and become clearer on structured problems like the Schelling model, underscoring the need to compare multiple surrogates rather than rely on a single one. Based on these findings, we recommend: (i) considering only well-adapted models given the problem and data at hand, (ii) preferring lightweight parametric models and using tree ensembles only when strong local predictive fidelity is required, or for assessing explanation stability via ensembles, and (iii) evaluating explanations jointly in terms of rank agreement, computational cost, accuracy, and robustness.

Our study opens new perspectives for research and future works should (i) validate these patterns on synthetic systems with known mechanisms (e.g., Lotka–Volterra or SIR dynamics) and with other post-hoc methods, (ii) compare multiple explainers to disentangle model versus explainer effects, (iii) incorporate human evaluation, (iv) develop automated **explanation-aware** model recommendation systems (AutoXAI) that jointly optimize accuracy, cost, and stability, and (v) generalize these findings and methods for time-series or domain-specific problems. Complementary work such as EXPO [Plumb *et al.*, 2020] shows that training-time regularization can make models more explanation-friendly. Our results complement this by demonstrating that **ML model choice alone** can amplify or mitigate explanation variability, and that strong inter-model agreement can serve as a practical signal of explanation robustness. In summary, no single model provides a uniquely correct explanation. Explanation interchangeability serves as a practical diagnostic: it reveals when inexpensive surrogate models produce explanations similar to those of complex models, and when explanations are unstable and require caution. Agreement between explanations reflects robustness across model choices under a fixed explainer, not access to ground-truth feature relevance.

References

[Anderson, 2016] Robert Anderson. The Rashomon effect and communication. *Canadian Journal of Communication*, 41(2):249–270, 2016.

[Apley and Zhu, 2020] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

[Beckh *et al.*, 2021] Katharina Beckh, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rueden. Explainable machine learning with prior knowledge: an overview. *arXiv preprint arXiv:2105.10172*, 2021.

[Bibal, 2020] Adrien Bibal. *Interpretability and Explainability in Machine Learning and Their Application to Nonlinear Dimensionality Reduction*. Doctoral thesis, University of Namur, Namur, Belgium, 2020.

[Breiman, 2001] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[Burges *et al.*, 2005] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[Caruana *et al.*, 2015] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[Christodoulou *et al.*, 2019] Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019.

[Cormack, 1971] Richard M Cormack. A review of classification. *Journal of the Royal Statistical Society: Series A (General)*, 134(3):321–353, 1971.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Garouani *et al.*, 2024] Moncef Garouani, Josiane Mothe, Ayah Barhrhouj, and Julien Aligon. Investigating the duality of interpretability and explainability in machine learning. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 861–867, 2024.

[Idrissi *et al.*, 2021] Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. Developments and applications of shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling & Software*, 143:105115, 2021.

[Järvelin and Kekäläinen, 2002] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[Kennedy and O'Hagan, 2001] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[Krishna *et al.*, 2022] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2022.

[Lakkaraju and Rudin, 2017] Himabindu Lakkaraju and Cynthia Rudin. Learning cost-effective and interpretable treatment regimes. In *Artificial intelligence and statistics*, pages 166–175. PMLR, 2017.

[Lattimore and Hutter, 2013] Tor Lattimore and Marcus Hutter. No free lunch versus occam's razor in supervised learning. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence: Papers from the Ray Solomonoff 85th Memorial Conference, Melbourne, VIC, Australia, November 30–December 2, 2011*, pages 223–235. Springer, 2013.

[Le *et al.*, 2023] Phuong Quynh Le, Meike Nauta, Shreyasi Pathak Van Bach Nguyen, Shreyasi Pathak, Jörg Schlötterer, and Christin Seifert. Benchmarking explainable ai-a survey on available toolkits and open challenges. In *IJCAI*, pages 6665–6673, 2023.

[Levy *et al.*, 2025] Jordan Levy, Clément Blanco-Volle, Nicolas Verstaevel, Benoit Gaudou, and Vincent Talon. TimeCIEL: contextual interactive ensemble learning for time series classification. In *23rd International Conference on Practical applications of Agents and Multi-Agent Systems (PAAMS 2025)*, 2025.

[Livet and Varenne, 2020] Pierre Livet and Franck Varenne. Artificial intelligence: philosophical and epistemological perspectives. In *A Guided Tour of Artificial Intelligence Research: Volume III: Interfaces and Applications of Artificial Intelligence*, pages 437–455. Springer, 2020.

[Löfström *et al.*, 2022] Helena Löfström, Karl Hammar, and Ulf Johansson. A meta survey of quality evaluation criteria in explanation methods. In *International Conference on Advanced Information Systems Engineering*, pages 55–63. Springer, 2022.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[Lundberg and Lee, 2019] Scott M. Lundberg and Su-In Lee. Shap (shapley additive explanations). https://github.com/shap/shap, 2019. GitHub repository.

[Mehdiyev *et al.*, 2025] Nijat Mehdiyev, Maxim Majlatow, and Peter Fettke. Interpretable and explainable machine learning methods for predictive process monitoring: A systematic literature review. *Artificial Intelligence Review*, 58(12):378, 2025.

[Minh *et al.*, 2022] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022.

[Moss *et al.*, 2022] Laura Moss, David Corsar, Martin Shaw, Ian Piper, and Christopher Hawthorne. Demystifying the black box: the importance of interpretability of predictive models in neurocritical care. *Neurocritical care*, 37(Suppl 2):185–191, 2022.

[Müller *et al.*, 2023] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the Rashomon effect in explainable machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 462–478. Springer, 2023.

[Olson *et al.*, 2017] Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10(1):36, 2017.

[Palar *et al.*, 2025] Pramudita Satria Palar, Paul Saves, Muhammad Daffa Robani, Nicolas Verstaevel, Moncef Garouani, Julien Aligon, Koji Shimoyama, Joseph Morlier, and Benoit Gaudou. Interpretable and explainable surrogate modeling for simulations: A state-of-the-art survey and perspectives on explainable AI for decision-making. *ArXiV preprint*, 2025.

[Peterson *et al.*, 2024] Ryan A Peterson, Max McGrath, and Joseph E Cavanaugh. Can a transparent machine learning algorithm predict better than its black box counterparts? a benchmarking study using 110 data sets. *Entropy*, 26(9):746, 2024.

[Plumb *et al.*, 2020] Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems*, 33:10526–10536, 2020.

[Retzlaff *et al.*, 2024] Carl O Retzlaff, Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Roettger, Heimo Mueller, and Andreas Holzinger. Post-hoc vs ante-hoc explanations: XAI design guidelines for data scientists. *Cognitive Systems Research*, 86:101243, 2024.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[Saves *et al.*, 2024] P. Saves, R. Lafage, N. Bartoli, Y. Diouane, J. H. Bussemaker, T. Lefebvre, J. T. Hwang, J. Morlier, and J. R. R. A. Martins. Smt 2.0: A surrogate modeling toolbox with a focus on hierarchical and mixed variables gaussian processes. *Advances in Engineering Software*, 188:103571, 2024.

[Schelling, 1969] Thomas C. Schelling. Models of segregation. *The American Economic Review*, 59(2):488–493, 1969.

[Smilkov *et al.*, 2017] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Wang *et al.*, 2023] Haomiao Wang, Emmanuel Doumard, Chantal Soulé-Dupuy, Philippe Kemoun, Julien Aligon, and Paul Monsarrat. Explanations as a new metric for feature selection: a systematic approach. *IEEE Journal of Biomedical and Health Informatics*, 27(8):4131–4142, 2023.

[Wilhelm and Zweig, 2024] Alexander Wilhelm and Katharina A Zweig. Hacking a surrogate model approach to XAI. *arXiv preprint arXiv:2406.16626*, 2024.