Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Are Feature Selection and Feature Attribution the Same? A Comparative Survey -- Extended Version

Muhammad Rajabinasab

rajabinasab@imada.sdu.dk

University of Southern Denmark https://orcid.org/0009-0006-7045-3998

Arthur Zimek

zimek@imada.sdu.dk

University of Southern Denmark https://orcid.org/0000-0001-7713-4208

Research Article

Keywords: Feature Selection, Feature Attribution, Explainable AI, Comparative Analysis

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10855689/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Are Feature Selection and Feature Attribution the Same? A Comparative Survey – Extended Version

MUHAMMAD RAJABINASAB and ARTHUR ZIMEK*, University of Southern Denmark, Denmark

Feature Attribution methods in Explainable AI (XAI) such as Shapley Values (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) provide a Feature Importance value and thus resemble classic Feature Selection methods which calculate a score and form a ranking for features. Despite the fact that these methods belong to different fields of research, their technical similarities seem to be more than their discrepancies. Hence, we study in this paper the question: Are feature selection and feature attribution methods the same? We present a survey on the theoretical background and the design choices of feature selection and feature attribution methods. We provide a comprehensive unified evaluation framework to evaluate and compare feature selection and feature attribution methods using a combined set of metrics. We also provide studies and analyses in different settings to investigate different properties of feature selection and feature attribution methods. With this comparative survey, we aim to shed light on the different aspects of feature selection and feature attribution methods, their similarities, differences, and unique strengths and weaknesses, bridging the gap between the two groups of methods and bringing the fields together.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Feature selection**; *Machine learning*; *Artificial intelligence*; • **Information systems** → Data mining.

Additional Key Words and Phrases: Feature Selection, Feature Attribution, Explainable AI, Comparative Analysis

1 Introduction

In the landscape of Data Mining and Machine Learning, a critical distinction exists between Feature Selection and Feature Attribution, despite the fact that they represent the two sides of the same analytical coin. Feature selection is the more traditional and classic approach to find a subset of relevant and informative variables within a raw dataset to use in a downstream machine learning or data mining task [23]. Imagine trying to predict the success of a food cooked by different chefs. Clearly, the quality of the raw material and factors like the right temperature and heating duration are more important than the color of the restaurant's manager's car, or the chef's favorite movie. On the other hand, feature attribution aims to find out which features were actually used by the predictive model to predict the success of the food. In this case, hypothetically, there might be a chance that, based on the distribution of the data, its variables, and intrinsic properties of the predictive model, a chef whose favorite movie is *Cinema Paradiso* would *spuriously appear* to be significantly more likely to deliver success in cooking, therefore highlighting *the chef's favorite movie* as the most important feature.

From a philosophical perspective, the idea behind feature selection roots in the principle of parsimony, or Occam's Razor: the idea that the simplest explanation—the one with the fewest moving parts—is preferred [8]. According to statisticians like Pearson and Fisher [20], including irrelevant information does not just add noise, it actively obscures the truth, leading to the phenomenon known as *the curse of dimensionality* [7, 27, 39, 58]. Feature *selection* aims to answer *"What do I need to know?"*, whereas feature *attribution* proposes another perspective raising the question: *"What did the model actually use?"* One might consider feature attribution as a continuation of more sophisticated wrapper-based feature selection methods [25] where the feature selection process focuses on the signals that truly drive the outcomes. This highlights a philosophical shift from the pursuit of *objective truth* to the pursuit of *accountability*. If feature selection is guided by the ancient principle of parsimony—an attempt to simplify the world—feature attribution

Authors' Contact Information: MUHAMMAD RAJABINASAB, rajabinasab@imada.sdu.dk; ARTHUR ZIMEK, zimek@imada.sdu.dk, University of Southern Denmark, Odense, Denmark.

is guided by the modern principle of *right to explanation* [21]. It is deeply rooted in epistemology, specifically the study of how we justify our beliefs. We no longer demand simplicity of the model, we demand it to be interrogable [34].

Despite the fact that feature selection and feature attribution seem to be deeply connected to one another, their evaluation criteria are quite different. Feature selection methods mostly go through supervised and unsupervised model-dependent evaluation [29, 38], although there have been attempts to conduct model-agnostic evaluation as well [40]. There are also attempts to investigate intrinsic aspects of the feature selection methods, such as stability [36]. Feature attribution methods, however, are model-dependent in nature. Despite the fact that Explainable AI (XAI) methods such as Shapley Values (SHAP) [32] and Local Interpretable Model-Agnostic Explanations (LIME) [44] are considered model-agnostic, the feature importance values they yield are still connected to a predictive model's decision-making process. Feature attribution methods are often evaluated based on criteria such as faithfulness and stability [5], which treat feature importance values as an attempt towards explainability and transparency, rather than simplicity.

1.1 Motivation

To the best of our knowledge, no previous study exists which provides a comprehensive comparative study of feature selection and feature attribution methods. Most well-known surveys on *feature selection* provide a general perspective of the advances in feature selection [11, 33], some surveys offer specific perspectives, such as data [29]. Other attempts focus on specific subgroups of feature selection methods, such as evolutionary algorithms [55]. Some surveys only focus on a specific task such as classification [51], others offer a wide range of downstream tasks in their study [18]. Also *feature attribution*, or in general XAI methods, gained a lot of attention in the recent years due to their applicability in providing trust and transparency in different AI-based systems. There exist comprehensive surveys on XAI methods [2, 16]. There are also more specialized surveys, for instance on argumentative XAI [15], or on our topic of interest, feature attribution [1]. There are also application-based surveys, for instance on medical AI [53] or cyber security [57].

Despite attempts to use XAI for developing feature selection methods [54, 56], to the best of our knowledge, there exists no work offering a comprehensive theoretical or empirical comparison of these two groups of methods. The existing gap between these two groups motivates the conduction of a comprehensive study to discuss and compare most important members of both groups, both theoretically and empirically, to shed light on their similarities and differences, and on opportunities to transfer the advances made in one group to the other.

1.2 Research Questions

Is the only difference between feature selection and feature attribution methods the philosophy behind their design and development, or are there more fundamental differences? This overall guiding question is detailed in the following research questions:

Research Question 1. *What are the mechanisms behind the fundamental feature selection and attribution methods?*

Research Question 2. *What are the evaluation criteria for feature selection and for feature attribution methods?*

Research Question 3. *From a variety of aspects such as philosophy, design, and evaluation criteria, are feature selection and attribution different or the same? And if they are different, can a criterion from one group be applied to the other?*

Research Question 4. *How do different methods from different groups compare to each other in different aspects? Is there a clear pattern for discrimination between feature selection and feature attribution?*

1.3 Contributions

In this paper, we bridge the gap between feature selection and feature attribution methods by conducting a comparative survey between the two groups of methods and their evaluation criteria. The main contributions are:

- reviewing the fundamental feature selection and feature attribution methods;
- discussing details of the most important and influential evaluation criteria for feature selection and for feature attribution techniques;
- a theoretical comparison between the two groups of methods;
- investigating the evaluation criteria for each group, their similarities and differences;
- providing a comprehensive comparative analysis of the methods based on all reviewed evaluation criteria and a variety of benchmark datasets to determine the best methods, understand performance patterns, and study the applicability of the evaluation criteria of one group to the other.

1.4 Outline

The remainder of this article is structured as follows: in Section 2, we provide a review of fundamental feature selection and feature attribution methods. In Section 3, we provide a review of the evaluation criteria employed to evaluate the performance of feature selection and feature attribution methods. In Section 4, we detail the experimental setup. In Section 5, we present the results of the comparative analysis, highlighting the performance of a selection of methods based on a comprehensive list of evaluation criteria. In Section 6, we provide a theoretical discussion to shed light on the causes and effects behind the observed experimental results. Finally in Section 7, we conclude the comparative survey by providing a summary, a list of limitations, and suggestions for future research.

2 Feature Selection and Feature Attribution Methods

Feature selection and feature attribution methods, as discussed, conduct the same task from two different perspectives. In feature selection, the goal is selecting the most informative features globally. For feature attribution, the aim is attributing the model's output to specific inputs. While feature selection determines which variables should be considered for inclusion in a model's architecture, feature attribution determines which of those variables were most influential for a specific conclusion. This section aims to answer Research Question 1.

Philosophically speaking, feature selection methods were designed as an attempt towards simplification, a step in the direction of addressing the curse of dimensionality by removing redundant or least important features before conducting a downstream machine learning or data mining task. Feature attribution on the other hand, is an attempt towards explainability, and it comes in after a machine learning or data mining task has already been performed. However, some feature selection methods (so-called wrapper methods) actually use a machine learning model in order to decide which features have the least effect on a model's performance. If these techniques are still categorized as feature selection, should the same principle not be applied to feature attribution techniques? In order to answer this question we need to provide theoretical and empirical details. In this section, we first provide a taxonomy of the feature selection methods to understand their different families better. Then, we provide a review of fundamental feature selection and feature attribution methods, with a focus on the underlying mechanisms of these methods. In order to make the review more coherent, we focus on supervised feature selection methods.

2.1 Taxonomy of Feature Selection Methods

Before detailing individual algorithms, it is useful to situate them within the broader taxonomy of feature selection strategies, which are conventionally divided into filter, wrapper, and embedded methods [11, 23].

Filter methods evaluate features using intrinsic statistical properties of the data, such as correlation, mutual information, or variance, independently of any downstream learning algorithm. Because the evaluation criterion is decoupled from the predictor, filter methods are computationally inexpensive and scale well to high-dimensional data, but they risk ignoring feature dependencies and interactions that only become apparent once a specific model is fit [23].

Wrapper methods, by contrast, treat the learning algorithm itself as a black-box and use its performance, typically measured through cross-validation, as the evaluation criterion for candidate feature subsets [25]. Wrapper methods conduct a search over the space of possible subsets, using strategies ranging from exhaustive and greedy sequential search to more sophisticated heuristics. Because they explicitly account for interactions between features and the induction algorithm, wrappers tend to yield subsets that are better tailored to the model at hand, at the cost of substantially higher computation due to repeated model retraining and a higher risk of overfitting the feature subset to the particular training sample [25].

Embedded methods occupy a middle ground: feature selection is performed as an integral part of the model training process, rather than as a separate preprocessing or wrapping step. LASSO is a canonical example, since the sparsity-inducing penalty performs selection during the optimization of the model's own objective function [52]. Embedded methods are generally more efficient than wrappers, since they avoid the repeated retraining cycles, while still being sensitive to feature interactions in a way that pure filter methods are not [11].

2.2 Supervised Feature Selection Methods

Supervised feature selection methods utilize the label information to identify subsets of features that maximize predictive utility, whether it is classification or regression. In the following, we detail for some of the most fundamental feature selection techniques how they carry out the feature selection task.

2.2.1 Mutual Information Maximization. Mutual Information (MI) is a filter-based feature selection method originated from Shannon's Information Theory, and was popularized as a feature selection criterion by Battiti [6]. It quantifies the amount of information obtained about the target variable Y by observing a feature X_j:

$$MI(X_j; Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}_j} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \tag{1}$$

where $\mathcal{X}_j$ and $\mathcal{Y}$ denote the (discretized, if continuous) value domains of feature X_j and target Y, respectively, $p(x, y)$ is their joint probability mass function, and $p(x)$, $p(y)$ are the corresponding marginals. Unlike Pearson correlation [12], MI captures any statistical dependency, including non-linear relationships [13]. Operationally, features are ranked (or greedily selected one at a time) by the magnitude of $MI(X_j; Y)$, with greedy variants such as the method by Battiti [6] additionally penalizing redundancy with already-selected features rather than scoring each feature in isolation. In high-dimensional spaces, MI is often estimated using k-nearest neighbor methods to bypass the difficulty of density estimation for continuous variables; this estimation step is itself a modeling choice, since it reintroduces smoothness or locality assumptions that MI, as a criterion, was originally meant to avoid. It serves as a fundamental benchmark for identifying impactful and non-redundant information.

2.2.2 Least Absolute Shrinkage and Selection Operator (LASSO). LASSO is an embedded method that performs feature selection via L_1 regularization. Given n training samples, each with a target $y_i \in \mathbb{R}$ and a d-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^d$, LASSO adds the sum of absolute values of the coefficients $\beta = (\beta_1, \ldots, \beta_d)$ as a penalty term to the ordinary least squares loss function, shrinking the coefficients of non-informative features exactly to zero [52]:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{d} |\beta_j| \right\} \tag{2}$$

where $\hat{\beta}$ denotes the resulting estimated coefficient vector and $\lambda \geq 0$ is the tuning (regularization) parameter controlling the strength of the penalty. Mechanically, this exact zeroing is a consequence of the geometry of the L_1 constraint region: its corners lie on the coordinate axes, so the optimum of the loss function is disproportionately likely to land exactly on an axis (i.e., with one or more coefficients at zero), unlike the smooth, rounded constraint region induced by an L_2 penalty, which shrinks coefficients toward zero without ever reaching it [52]. The tuning parameter λ controls the sparsity of the solution, allowing for a continuous transition from a full model ($\lambda \to 0$) to a null model ($\lambda \to \infty$). This is particularly effective for high-dimensional datasets where $d > n$.

2.2.3 Fisher Score. The Fisher Score seeks features that minimize the within-class variance while maximizing the between-class variance, a criterion rooted in Fisher's original linear discriminant [20] and later formalized for feature ranking as:

$$FS(X_j) = \frac{\sum_{k=1}^{c} n_k (\mu_{kj} - \mu_j)^2}{\sum_{k=1}^{c} n_k \sigma_{kj}^2}, \tag{3}$$

where c is the total number of classes, n_k is the number of samples belonging to class k (so that $n = \sum_{k=1}^{c} n_k$ is the total sample count), μ_{kj} and σ_{kj} represent the mean and standard deviation, respectively, of feature j within class k, and μ_j is the overall (class-independent) mean of feature j across all n samples. Concretely, the score is computed independently, per feature, by comparing how far apart the class means are from the overall mean (the numerator) against how spread out each class is around its own mean (the denominator); features are then ranked and the top-scoring features are retained. It is computationally efficient and independent of the intrinsic properties of the classifier used. A feature with a high Fisher Score indicates that the class distributions are well-separated in that feature's dimension [22].

2.2.4 Recursive Feature Elimination. Recursive feature elimination (RFE), introduced by Guyon et al. [24] in the context of gene selection with support vector machines, is a greedy wrapper-based approach that iteratively removes features based on their direct impact on the performance of the downstream task. Concretely, at each iteration the model is trained on the current feature set, each remaining feature j is scored by a model-specific importance criterion (in the original formulation, the squared SVM weight w_j^2, quantifying each feature's contribution to the margin), and the lowest-scoring feature (or a small batch of them) is discarded before retraining; the process repeats until the desired number of features is reached. The recursive feature elimination is very effective in identifying most impactful features for the downstream task, and is superior to filter methods in capturing feature interactions, although it is computationally more expensive due to the iterative retraining of the model.

2.2.5 White-box Feature Selection. Some machine learning models allow white-box access to look directly into the model and the importance of each variable after the training is done. For example, Logistic Regression [14] provides the user with coefficients, Decision Trees [10] are based on split statistics (e.g., Gini impurity or information-gain reduction

attributable to a feature), and Random Forests [9] show a decrease in the mean impurity when aggregated across many trees. This clearly can be an effective method for feature selection with respect to the model's preferences.

2.2.6 ReliefF Algorithm. ReliefF is an extension of the original Relief algorithm [26], designed for multi-class problems and noisy data. It is an instance-based method that estimates the quality of features by how well their values distinguish between instances that are near to each other [26]:

$$W[j] = W[j] - \sum_{k=1}^{m} \frac{diff(j, R, H_k)}{m \cdot k} + \sum_{C \neq class(R)} \frac{P(C)}{1 - P(class(R))} \sum_{k=1}^{m} \frac{diff(j, R, M_k(C))}{m \cdot k}, \tag{4}$$

where $W[j]$ is the running relevance weight accumulated for feature j, R is a training instance sampled at random at each iteration, m is the number of nearest neighbors considered per class (a hyperparameter distinct from the total sample count n), H_k are the k-th nearest "hits" (neighbors of R sharing its class), $M_k(C)$ are the k-th nearest "misses" from a different class C, $class(R)$ denotes the class label of R, $P(C)$ is the prior probability of class C in the training data, and $diff(j, A, B)$ is a distance function on feature j between two instances A and B (e.g., normalized absolute difference for numeric features, or a mismatch indicator for categorical ones). Operationally, the algorithm repeatedly samples an instance R, locates its nearest same-class neighbors (hits) and nearest neighbors from each other class (misses), and updates each feature's running weight $W[j]$ according to whether that feature's value differs more from the misses than from the hits; a feature that consistently separates classes at the local instance level accumulates a high weight. This allows for the detection of features that are highly conditional on other features.

2.2.7 Summary. The discussed methods belong to the different branches of the taxonomy. Mutual information maximization and the Fisher Score are filter methods, LASSO is an embedded method, recursive feature elimination is a wrapper method, and white-box feature selection and ReliefF span aspects of both filter and embedded, depending on how directly they draw on the induction algorithm's internal structure. As observed in the details of each of these algorithms, there is an operational aspect to the feature selection process concerning how candidate subsets are constructed. Forward methods build a subset incrementally from the empty set; backward methods, recursive feature elimination being the paradigmatic example, start from the full feature set and remove features one at a time, retraining or re-scoring after each removal [24, 25, 42]. The majority of the feature selection methods however, evaluate the entire feature set simultaneously, providing feature importance scores with no notion of sequentiality.

One commonality between all these algorithms is their global view of the feature selection process. For these methods, the features are meaningful and hence considered through a global lens of the data. The effect of a feature is investigated across the entire dataset. This global point of view is perhaps the main characteristic that we should keep in mind for the comparison with feature attribution methods.

2.3 Taxonomy of Feature Attribution Methods

Feature attribution methods can be organized along a two-dimensional taxonomy defined by their underlying mathematical mechanism and their required level of model access. Mechanistically, these techniques generally fall into three distinct branches: *surrogate approximations*, which fit simple local models to approximate local decision boundaries; *perturbation and variance-based frameworks*, which measure output shifts by systematically altering or marginalizing input features; and *structural path-based methods*, which exploit internal model properties directly. Operationally, methods can be differentiated according to the degree of required model transparency. The surrogate and perturbation branches form a *model-agnostic* tier that treats the target model as a complete black-box, requiring only input–output

query access to compute attributions. In contrast, path-based approaches constitute a *model-specific* tier that necessitates white-box access to internal architecture, relying on direct inspection of components such as decision tree graphs or continuous gradient fields.

2.4 Feature Attribution Methods

Feature attribution focuses on local interpretability, assigning a contribution value to each input feature for a single prediction. To ensure broad applicability, the methods selected for this survey are predominantly model-agnostic. Throughout this subsection, $x = (x_1, \ldots, x_d)$ denotes the instance being explained, d the total number of input features, n (where relevant) the number of samples in a dataset, and f (or F, where a differentiable model is assumed) the predictive model being explained.

2.4.1 SHapley Additive exPlanations (SHAP). SHAP provides a unified framework for interpreting any model by using Shapley values, originating from cooperative game theory [32, 47]. It is a method that satisfies the axioms of local accuracy, missingness, and consistency. The attribution ϕ_i is the weighted average of the marginal contribution of feature i across all possible feature subsets S [32]:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} [f_x(S \cup \{i\}) - f_x(S)] \tag{5}$$

where $N = \{1, \ldots, d\}$ is the full set of feature indices, S ranges over all subsets ("coalitions") of N that exclude feature i, $|S|$ is the size of a given coalition, and $f_x(S)$ denotes the model's expected prediction for instance x when only the features in S are known (features outside S are marginalized out, e.g., by averaging over the training distribution). Mechanically, this means enumerating every possible coalition of other features S, checking how much the prediction changes when feature i joins that specific coalition, and averaging that marginal change over all coalitions, weighted so that every possible ordering of features is equally likely; SHAP's axioms guarantee that this is the unique attribution scheme with that fairness property, which is why the exact computation is exponential in d and practically infeasible beyond a small number of features. SHAP values provide a contrastive explanation, comparing the prediction to the average output across the training set.

2.4.2 LIME (Local Interpretable Model-agnostic Explanations). LIME approximates the model f locally around a specific point x using an interpretable surrogate model g, drawn from a class G of simple, human-readable models (e.g., sparse linear models or shallow decision trees) [45]. The surrogate model is trained on a set of perturbed samples weighted by their proximity to the original input:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{6}$$

where $\xi(x)$ is the resulting explanation (the fitted surrogate g, or equivalently its coefficients), $\mathcal{L}(f, g, \pi_x)$ is a fidelity loss measuring how poorly g approximates f on samples drawn near x, π_x is a proximity (kernel) weighting function that gives more weight to perturbed samples closer to x, and $\Omega(g)$ is a complexity penalty (e.g., the number of non-zero coefficients) that keeps g interpretable. Concretely, LIME generates a synthetic neighborhood of n perturbed samples around x by randomly altering its d feature values, queries the black-box model f for a prediction on each perturbed sample, and fits g to these input–output pairs, weighting each sample by π_x so that points closer to x matter more. LIME is fundamentally model-agnostic, treating the original model as a black-box and only requiring the ability to probe it with new inputs. The explanation is based on the simplified surrogate model fitted on the perturbed samples in the proximity of the query point.

2.4.3 Permutation Feature Attribution. The permutation method, originally proposed as a global importance measure for random forests [9] and later formalized in a model-agnostic way [19], calculates the importance of a feature by measuring the change in the model's prediction when the value of that feature is permuted or replaced by a baseline. For a local instance $x = (x_1, \ldots, x_d)$, we replace x_i with a value drawn from that feature's marginal distribution and observe the output shift [19]:

$$Attr_i(x) = f(x) - \mathbb{E}_{x'_i \sim P(X_i)}[f(x_1, \ldots, x'_i, \ldots, x_d)] \tag{7}$$

where $P(X_i)$ is the marginal (empirical) distribution of feature i over the training set, and $x'_i \sim P(X_i)$ denotes a value resampled from that distribution to replace x_i while holding all other $d-1$ features fixed. This reflects the "uniqueness" of the information provided by the feature for that specific prediction.

2.4.4 Feature Ablation. Feature ablation is another perturbation-based attribution method that evaluates a feature's impact by observing the model's response to its absence. For an instance $x = (x_1, \ldots, x_d)$, ablation mutes feature j by replacing it with a neutral baseline value b_j (e.g., median or zero) [49]. The attribution $a_j(x)$ is defined as the absolute shift in the output probability:

$$a_j(x) = \left| f(x_1, \ldots, x_j, \ldots, x_d) - f(x_1, \ldots, b_j, \ldots, x_d) \right| \tag{8}$$

Given a dataset of n instances, with $x^{(i)}$ denoting the i-th instance and $x^{(i)}_{j \leftarrow b_j}$ that same instance with feature j replaced by the baseline b_j, the global importance I_j is the mean absolute deviation across all n samples:

$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| f\left(x^{(i)}\right) - f\left(x^{(i)}_{j \leftarrow b_j}\right) \right| \tag{9}$$

2.4.5 Kernel SHAP. Kernel SHAP is a model-agnostic method to estimate SHAP values using a specifically weighted linear surrogate model [32]. It bridges the gap between LIME and Shapley values by showing that the Shapley values can be recovered through a weighted least squares regression, where each sampled coalition is represented as a binary vector $z' \in \{0,1\}^d$ (with a 1 in position j indicating that feature j is "present" in that coalition) and weighted by the SHAP kernel:

$$\pi_{x'}(z') = \frac{(d-1)}{\binom{d}{|z'|} |z'| (d - |z'|)}, \tag{10}$$

where d is the total number of features (as defined at the start of this subsection) and $|z'|$ is the number of "present" features in the coalition z'. In practice, this means sampling a manageable number of feature coalitions z' (rather than all 2^d of them), evaluating the model on each, and fitting a weighted linear regression whose weights $\pi_{x'}(z')$ are set precisely so that the regression coefficients converge to the exact Shapley values. Kernel SHAP therefore trades the combinatorial explosion of exact Shapley computation for a sampling-and-regression approximation. This allows for the calculation of theoretically grounded SHAP values without requiring model-specific internal access.

2.4.6 Sobol Indices (Global Sensitivity Analysis). Sobol indices, introduced by Sobol [48], are variance-based sensitivity measures that decompose the variance of the model output $Y = f(\mathbf{X})$, with $\mathbf{X} = (X_1, \ldots, X_d)$ the vector of (random) model inputs, into fractions that can be attributed to sets of input variables. While traditionally global, they are adapted for attribution by looking at local variance contributions:

$$S_i = \frac{V_i}{V(Y)} = \frac{V_{X_i}\left(\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i)\right)}{V(Y)} \tag{11}$$

where $V(Y)$ is the total variance of the model output over the input distribution, $\mathbf{X}_{\sim i}$ denotes all $d-1$ input variables except X_i, $\mathbb{E}_{\mathbf{X}_{\sim i}}(Y \mid X_i)$ is the expected output when X_i is held fixed and every other input is averaged out, $V_{X_i}(\cdot)$ is the variance of that conditional expectation taken over the possible values of X_i, and V_i is shorthand for this quantity (the first-order variance contribution of X_i). In practice, S_i is estimated by Monte Carlo sampling: repeatedly fixing X_i and randomizing every other input, then measuring how much of the total output variance is explained by that fixing: a large S_i means that pinning down X_i alone already accounts for most of the model's output variability. This method provides a rigorous mathematical framework for understanding the main effects and interaction effects of features without assuming model linearity or monotonicity.

2.4.7 Saabas (Tree Interpreter). The Saabas method [46] is a path-based feature attribution technique specifically tailored for tree-based models. It decomposes a model's prediction into a sum of feature contributions by calculating the change in the expected value of the target variable at each split in the decision tree. For a given instance x, the attribution ϕ_i for feature i is defined as the total change in the local mean value v across all nodes k where a split occurs on that feature:

$$Attr_i(x) = \sum_{k \in \text{path}(x)} \mathbb{1}(\text{split}_k = i) \cdot (v_{\text{child}(k,x)} - v_k), \tag{12}$$

where $\text{path}(x)$ is the ordered sequence of nodes from the tree's root to the leaf reached by instance x, split_k is the feature on which node k makes its split, $\mathbb{1}(\cdot)$ is the indicator function (equal to 1 if its argument is true, 0 otherwise), v_k represents the mean response of the training samples falling into node k, and $\text{child}(k,x)$ is the subsequent node in the decision path for instance x. Concretely, the method walks the single decision path that instance x takes from root to leaf and, at every split encountered, credits (or debits) the splitting feature with the resulting change in the node's mean prediction; for ensemble methods such as Random Forests, the final attribution is computed by averaging the contributions across all constituent trees. This path-dependence is also its main limitation relative to Tree SHAP: because only the single path taken by x is considered, the attribution can depend on the arbitrary order in which splits happen to occur, whereas Tree SHAP averages over all orderings via the Shapley formulation. While Saabas provides a computationally efficient decomposition of the prediction path, it is often seen as a heuristic precursor to the more axiomatically rigorous Tree SHAP [31].

2.4.8 Integrated Gradients (IG). Integrated Gradients [50] is a gradient-based attribution method that satisfies the "Completeness" axiom. The "Completeness" axiom states that the sum of the feature attributions for a given input equals the difference between the model's output for that input and its output for a baseline input. Assuming a differentiable model F, it addresses the issue of "flat" gradients in deep networks by integrating the gradient of the output with respect to the input along a straight path from a baseline x' to the input x:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{13}$$

where x' is a d-dimensional baseline input representing an absence of information (e.g., an all-zero or all-black image), and $\alpha \in [0,1]$ is an interpolation parameter that linearly traces the straight-line path between x' (at $\alpha = 0$) and x (at $\alpha = 1$), so that $x' + \alpha(x - x')$ is the point on that path at position α. In practice, this integral is approximated by a Riemann sum: a small number of intermediate points along the straight-line path from x' to x are constructed by discretizing α, the gradient of the model's output is computed at each interpolated point, and these gradients are averaged and multiplied by the total displacement $(x_i - x_i')$. Accumulating the gradient signal along the whole path avoids the problem of a single gradient evaluation at x being saturated (near-zero) even though the feature clearly

mattered to reach the prediction from the baseline. While gradient-based, this method is considered generalizable for any differentiable model, including neural networks, logistic regressions, and differentiable tree approximations.

2.4.9 Summary. To cover a broad spectrum of interpretability approaches, the feature attribution techniques analyzed here are selected as representatives from three main methodological branches, varying also in their required level of model access. From the *surrogate approximation* branch, LIME and Kernel SHAP are representative model-agnostic approaches that fit local interpretable surrogates via black-box probing. From the *perturbation and variance-based* branch, SHAP, Permutation/Ablation and Sobol Indices represent model-agnostic methods that evaluate feature impact through input substitution or coalition sampling. Finally, from the *structural path-based* branch, Saabas and Integrated Gradients are white-box representatives that leverage internal model mechanics, specifically tree decision paths and differentiable gradients.

Apart from the aforementioned distinctions, feature attribution methods have in common the local perspective of the feature importance calculation task. For the feature attribution methods, the importance of the features is evaluated per query. For instance in SHAP [47], a prediction is first made, and then the influence of each feature and its magnitude and direction of contribution to the prediction made is calculated. This might be the most common characteristic shared between all the feature attribution methods, and this characteristic is pointing exactly in the opposite direction of the philosophical lens feature selection methods make use of. It is noteworthy that this local approach can easily be aggregated via for instance, averaging, into a global view, which is also often done for these methods to derive feature importance for a whole dataset, but it is still an aggregation of local observations of feature importance.

2.5 Discussion

Having reviewed the mechanics of both families of methods, we can now return to the philosophical distinction, parsimony versus interrogability, "what do I need to know" versus "what did the model actually use", and test it against what the methods themselves actually do.

The operational core. Despite their different vocabularies, several feature selection and feature attribution methods share an identical operational core: remove or alter a feature, and observe the resulting change in output. Recursive feature elimination discards the lowest-scoring feature and retrains to observe the change in *aggregate model performance*; permutation and ablation-based attribution perturb or remove a feature and observe the change in a *single prediction*. The mechanism is the same perturb-and-observe logic; what differs is the unit of observation, dataset-level performance for the wrapper, instance-level output for the attribution method. This illustrates that feature attribution can be read as a continuation of wrapper-based feature selection: the wrapper's *remove-and-retrain* loop is, at the level of a single instance, structurally identical to the ablation-based attribution's *remove-and-re-observe* loop. It also directly addresses Research Question 3: a wrapper's evaluation criterion (change in performance upon removal) is transferable to the attribution setting almost unchangingly, simply by shrinking its scope from the dataset to the instance.

Connections to the taxonomy. The filter/wrapper/embedded taxonomy does not map cleanly onto feature attribution methods. Kernel SHAP requires access to the model to be queried at sampled coalitions, a wrapper-like dependency, yet the coefficients it recovers are dictated analytically by the Shapley axioms rather than discovered through search, which is filter-like in spirit. Saabas and Tree SHAP require white-box access to a tree's internal split structure, placing them alongside white-box feature selection, yet they operate squarely at the attribution, not the selection level. This is not merely a labeling inconvenience: it suggests that the filter/wrapper/embedded distinction encodes something specific

to *selection* decisions (how much of the model's search process is used to decide whether a feature stays) that does not have a clean analogue in *attribution* decisions (how much of the model's structure is used to decide how much a feature mattered). This is itself a partial answer to Research Question 3: the two fields are not simply the same taxonomy applied at different granularities.

Parsimony versus interrogability. We discussed that feature selection pursues objective simplification (Occam's razor) while feature attribution pursues accountability toward a human recipient (the right to explanation). The methods reviewed here largely bear this out, but with an important asymmetry. Mutual information maximization and the Fisher Score are audience-free: their scoring criteria make no reference to whether the resulting subset would be intelligible or convincing to anyone. SHAP's axioms (local accuracy, missingness, consistency), by contrast, are explicitly chosen because they make the resulting attribution *defensible* to a human recipient, not merely because they are mathematically minimal; this is a design choice, not an incidental property, and it is what separates SHAP's axiomatic program from an equally rigorous but audience-indifferent criterion like a Sobol index. The asymmetry sharpens our answer to Research Question 3: some evaluation criteria (perturb-and-observe) transfer easily across the feature selection and feature attribution boundary, while others (axiomatic fairness toward an explanation's recipient) are specific to attribution and have no obvious feature selection counterpart, since no one has historically asked a filter method to be "convincing."

Local versus global. Feature selection and feature attribution methods highlight two different philosophical perspectives of feature importance. Feature selection methods offer a holistic view, often looking at all of the features and instances in the dataset together. In other words, they look at the dataset as an entirety. On the other hand, feature attribution methods offer a local perspective; they consider importance as a notion of contribution of each feature to each individual prediction made by a model. This local perspective, however, can be aggregated (e.g., via averaging) to provide a holistic dataset-wise perspective of feature importance, while not all feature selection methods could work with a local perspective (although, as mentioned above, a wrapper's global evaluation criterion could be given a local focus and hence made local).

The curse of dimensionality. Finally, the two families relate to the curse of dimensionality in opposite directions. Feature selection methods exist to fight the curse *before* modeling, by removing features prior to training. Perturbation-based attribution methods, in exchange for their local fidelity, reintroduce a version of the same curse *after* modeling, at explanation time: exact SHAP values require summing over 2^d coalitions, and Sobol indices require variance estimates over an exponentially growing space of input combinations. Kernel SHAP's sampling-and-regression approximation exists precisely to dodge this explosion, in effect re-fighting, at explanation time, the same combinatorial battle that feature selection methods fought at training time. This symmetry suggests that dimensionality is not a problem specific to model training that explanation methods are exempt from; it resurfaces wherever a method must reason over combinations of features, regardless of which side of the training/explanation boundary it sits on. However, the curse of dimensionality is a motivation for both fields, making a dataset more manageable for learning or more interpretable after learning, by selecting some subset of dimensions to focus on.

2.6 Summary

In this section we reviewed some of the most fundamental methods for both feature selection and feature attribution that form the basis of the two respective research fields, situating each within a filter/wrapper/embedded taxonomy and detailing the operational mechanism, not merely the objective, behind each method's decisions. In summary, the main

difference is that feature attribution methods strictly need a predictive model, and they start providing insights after the prediction has taken place. On the other hand, feature selection methods use a combination of predictive, statistical, or structural algorithms in order to provide feature scores and rankings. Both families of methods are providing valuable insights on the importance of the features, feature selection offers a more global and dataset-wide view, whereas feature attribution often focuses on local, point-wise insights; the discussion above further shows that this apparent difference partly dissolves at the level of operational mechanism, even where it persists at the level of design philosophy.

3 Evaluation Criteria

In this section, we focus on answering Research Question 2 by providing insights into these metrics and a brief but comprehensive review of them, both in the field of feature selection and feature attribution. Our focus will be on the differences, underlying similarities, and the limitations of applying a metric from one field to the other. With the emergence of XAI and the growing focus on the explainability aspect, new evaluation metrics and measures have been developed. Despite the existence of some fundamental differences, some of these new metrics and measures seem to be very similar in nature to those proposed for the evaluation of feature selection methods.

3.1 Feature Selection Evaluation

Feature selection methods are primarily evaluated based on their ability to enhance (or maintain) the performance of downstream machine learning tasks while simultaneously reducing the dimensionality. Evaluation methods for feature selection can be divided into two main categories: (i) performance assessment on downstream supervised or unsupervised tasks, typically using metrics computed for a fixed number of selected features, and (ii) measures that capture broader properties such as overall performance trends, model-agnostic quality, fitness gain, or stability.

3.1.1 Evaluation Based on Downstream Tasks. The most common evaluation approach involves measuring how feature selection affects the quality of a subsequent machine learning task, most frequently classification (supervised) or clustering (unsupervised). These evaluations are usually point-based, i.e., they report performance achieved with a specific number of selected features.

Supervised Evaluation. In supervised settings, the effectiveness of a feature selection method is typically quantified by training one or more classifiers on the reduced feature set and calculating standard classification performance measures.

Accuracy (ACC): Accuracy is defined as the proportion of correctly classified instances:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{14}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure useful in the presence of class imbalance:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{15}$$

with $\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and $\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$.

Area Under the ROC Curve (AUC): The area under the receiver operating characteristic curve quantifies a classifier's ability to discriminate between classes across all decision thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR} \tag{16}$$

where TPR is the true positive rate and FPR is the false positive rate. Values of AUC close to 1 indicate excellent discriminative power, while AUC = 0.5 corresponds to random guessing.

Supervised metrics are used widely for the evaluation of feature selection methods. However, their values are strongly dependent on the choice of classifier, hyperparameter settings, and the dataset characteristics.

Unsupervised Evaluation. In unsupervised feature selection, performance is frequently evaluated by applying a clustering algorithm (usually k-means) to the selected features and comparing the resulting clustering labels against ground-truth class labels using external validation measures. Although these procedures are commonly called "unsupervised evaluation" in the literature, they are in fact, albeit based on an unsupervised downstream task (clustering), supervised evaluation procedures overall as they are eventually resorting to class labels.

Clustering Accuracy (CLSACC): Clustering accuracy measures the agreement between predicted cluster labels and true class labels after finding an optimal label correspondence:

$$\text{CLACC} = \frac{1}{n} \sum_{i=1}^{n} \delta\big(c_i, \text{map}(g_i)\big) \tag{17}$$

where n is the number of instances, c_i and g_i are the predicted cluster and true class labels, respectively, for instance i, $\text{map}(\cdot)$ is the optimal assignment obtained via the Hungarian method [37], and $\delta(\cdot,\cdot)$ is the Kronecker delta function.

Normalized Mutual Information (NMI): Normalized mutual information quantifies the shared information between the predicted clustering C and the true partitioning indicated by the ground-truth labels L, normalized by their entropies:

$$\text{NMI} = \frac{\text{MI}(L, C)}{\max\big(\text{Entropy}(L), \text{Entropy}(C)\big)} \tag{18}$$

where $\text{MI}(\cdot,\cdot)$ denotes mutual information (Eq. 1). The measure ranges from 0 (independent) to 1 (perfect agreement).

These external clustering metrics, while informative, still rely on ground-truth labels, which is in contrast to the unsupervised nature of the downstream task. If we aim to evaluate an unsupervised feature selection algorithm on a dataset which does not have any labels, these so-called "unsupervised" evaluation procedures would be rendered ineffective as much as the supervised evaluation.

3.1.2 Specialized Evaluation Measures. Several specialized metrics have been proposed to overcome limitations of single-point and downstream task dependent evaluations by providing more holistic, model-agnostic, or robustness-oriented insights.

FSDEM Score. The Feature Selection Dynamic Evaluation Metric (FSDEM) [38] aggregates performance across varying numbers of selected features. Let d be the total number of features, $P(f)$ the value of a downstream performance measure obtained using exactly f selected features, and $g(x)$ a continuous function approximated based on the observed pairs $\big\{\big(f, P(f)\big)\big\}_{f=1}^{d}$. The FSDEM score is defined as the normalized area under this curve over a chosen interval $[a, b]$:

$$\text{FSDEM} = \frac{1}{b-a} \int_a^b g(x)\, dx. \tag{19}$$

When the underlying performance measure P is bounded in $[0, 1]$, so is FSDEM, with values closer to 1 indicating consistently good performance. FSDEM allows approximating the overall performance of the entire feature selection process based on a number of observations, leading to a more holistic overview compared to the common single-point evaluation.

Average Angle Difference (AAD). The Average Angle Difference [40], also formally denoted as $\overline{\Delta\theta}$, provides a model-agnostic assessment of feature selection quality by quantifying the average change in the direction of the eigenvector corresponding to the first principal component when non-selected features are removed. Let F be the set of selected features and F^C its complement. For each feature $f \in F^C$, construct the modified dataset $\mathbf{D}_{f^0}$ by setting all values of feature f to zero. The AAD is then

$$AAD = \frac{1}{|F^C|} \sum_{f \in F^C} \Delta\theta\big(\mathbf{D}, \mathbf{D}_{f^0}\big), \tag{20}$$

where $\Delta\theta$ denotes the angle difference between the eigenvectors corresponding to the first principal components. Smaller values of $\overline{\Delta\theta}$ (closer to 0) indicate that the removed features were largely redundant or uninformative, hence affecting the first and most important principal component the least.

Baseline Fitness Improvement (BFI). The Baseline Fitness Improvement [35] metric quantifies the net gain obtained by feature selection relative to using the complete feature set:

$$\mathrm{BFI}(s) = f(s) - f(s^*), \tag{21}$$

where $f(s)$ is the fitness of subset s, $f(s^*)$ is the baseline fitness of the full feature set, and the fitness function is typically expressed as

$$f(s) = k_c \cdot f_c(s) + k_p \cdot f_p(s), \tag{22}$$

with $f_c(s)$ a classifier performance term and $f_p(s)$ an exponentially growing penalty on the number of selected features. Positive BFI values reflect a favorable trade-off between predictive performance and dimensionality reduction.

Stability Measures. Stability quantifies the robustness of feature selections under different conditions such as the presence of noise, the perturbations in the input training data, and the performance gain.

Nogueira et al. Stability Metric: The stability metric proposed by Nogueira et al. [36] is a variance-based stability index based on Pearson's correlation coefficient. It reflects the stability of the feature selection task in hte presence of noise, defined as

$$\hat{\Phi}(Z) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)}, \tag{23}$$

where d is the total number of features, $\bar{k}$ is the average number of selected features, and s_f^2 is the selection variance of feature f. Interpretive thresholds are $\hat{\Phi} > 0.75$ (high stability), $0.40 \leq \hat{\Phi} \leq 0.75$ (intermediate), and $\hat{\Phi} < 0.40$ (poor stability).

Kuncheva Stability Index: For K independent feature selection runs yielding subsets $S_1(k), \ldots, S_K(k)$ of fixed size k, the stability according to Kuncheva [28] is

$$\mathcal{I}_S(A(k)) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} I_C\big(S_i(k), S_j(k)\big), \tag{24}$$

where I_C is the consistency index measuring subset overlap. Kuncheva Stability Index reflects the robustness and reproducibility of a feature selection algorithm under perturbations in the input training data.

FSDEM-Derived Stability Score: A complementary stability measure can be obtained from the first-order derivative of the fitted performance curve $g(x)$ of the FSDEM measure (Eq. 19):

$$\text{STAB} = \frac{1}{b-a+1} \sum_{x=a}^{b} g'(x). \tag{25}$$

Positive values close to 1 indicate that including additional selected features consistently improve (or at least preserve) downstream performance, reflecting informative rather than merely redundant selections [38].

3.1.3 Discussion. We reviewed evaluation metrics proposed and used for the evaluation of feature selection algorithms. A natural question that follows is whether any of these measures can also be applied to the evaluation of feature attribution methods. The downstream-task-based measures (ACC, F1-Score, AUC, CLSACC, NMI) are, in principle, transferable, since they only require a discrete subset of features to operate on; obtaining such a subset from a feature attribution method simply requires thresholding the attribution scores at a chosen budget k and retaining the top-k features, after which the same evaluation procedure used for feature selection applies unchanged. The same holds for AAD, which is model-agnostic and only relies on the notion of a retained/removed feature set rather than on how that set was produced. Once this discretization step is applied, however, all of the reviewed feature selection measures become usable for feature attribution methods as well, which is the approach we adopt in our own experimental comparison of the two families of methods. What remains missing, regardless of whether the underlying method is a feature selector or an attribution technique, is a measure capable of truly unsupervised evaluation: among the metrics discussed, only AAD [40] does not rely on labels, but being model and task-agnostic, it also does not directly reflect the downstream performance that both families of methods ultimately aim to support.

3.2 Feature Attribution Evaluation

The evaluation of feature attribution methods involves a point-based view on the evaluation, similar to most of the feature attribution methods, in contrast to the global, budget-indexed perspective adopted for feature selection. In this section, we survey the corresponding evaluation metrics and measures for feature attribution methods, covering the metrics included in the OpenXAI toolkit [4], before returning in the summary to the question, which of these measures could, in principle, be adapted for the evaluation of feature selection methods.

3.2.1 Model-Data Relationship Sanity Checks. A critical requirement for any supervised feature selection algorithm is its dependence on the relationship between the features X and the target labels Y. We discussed earlier that as we consider comparing feature attribution and feature selection, we are focused on supervised feature selection algorithms. To evaluate this dependency, we adapt the *Randomized Label Test*—originally proposed as a sanity check for saliency maps [3]—to the domain of feature selection and feature attribution.

Randomized Label Test (RLT). The Randomized Label Test assesses whether a feature selection method is sensitive to the underlying data-label relationship or if it relies solely on the internal structure of the input features (e.g., variance or marginal distributions).

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ be the original dataset and $\mathcal{S}(\mathcal{D})$ be the set of features selected by a given algorithm. We construct a permuted dataset $\mathcal{D}_{perm} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^{n}$, where $\tilde{y}$ is a random permutation of the original labels Y, effectively

breaking the relationship $P(Y|X)$. The sanity check is defined by the similarity between the two selections:

$$\text{Sim}(\mathcal{S}(\mathcal{D}), \mathcal{S}(\mathcal{D}_{perm})) \approx 0 \tag{26}$$

where $\text{Sim}(\cdot)$ is a similarity measure such as the Jaccard index. A feature selection method "passes" the sanity check if its output on $\mathcal{D}_{perm}$ is significantly different from its output on $\mathcal{D}$. If the selected features remain largely unchanged despite the randomization of labels, the method is deemed to be capturing data artifacts or unsupervised properties rather than predictive information, rendering it unreliable for supervised tasks.

3.2.2 OpenXAI Evaluation Framework. To provide a comprehensive perspective on feature attribution, we survey metrics from the OpenXAI toolkit [4], which categorizes evaluation into ground-truth faithfulness, predictive faithfulness, and stability.

Predictive Faithfulness. Predictive faithfulness measures how well the attribution corresponds to the model's actual decision-making process by observing the change in output when features are perturbed.

Prediction Gap on Important Features (PGI): PGI measures the difference in prediction probability resulting from perturbing features deemed influential by the explanation e_x:

$$\text{PGI}(x, f, e_x, k) = \mathbb{E}_{x' \sim \text{perturb}(x, e_x, \text{top-}k)} \left[|\hat{y} - f(x')| \right] \tag{27}$$

where $\hat{y}$ is the original prediction and x' is generated by adding Gaussian noise to the top-k features while holding others constant.

Prediction Gap on Unimportant Features (PGU): Conversely, PGU measures the change when perturbing features deemed unimportant:

$$\text{PGU}(x, f, e_x, k) = \mathbb{E}_{x' \sim \text{perturb}(x, e_x, \text{non top-}k)} \left[|\hat{y} - f(x')| \right] \tag{28}$$

For both PGI and PGU, the reported value is typically the Area Under the Curve (AUC) calculated over all possible values of k.

Stability Metrics. Stability quantifies the robustness of an explanation by measuring the maximum change in the attribution e_x relative to small changes in the input, representation, or output.

Relative Input Stability (RIS): RIS evaluates how much the explanation changes when the input x is slightly perturbed within a neighborhood N_x:

$$\text{RIS}(x, x', e_x, e_{x'}) = \max_{x'} \frac{\left\| \frac{e_x - e_{x'}}{e_x} \right\|_p}{\max\left(\left\| \frac{x - x'}{x} \right\|_p, \epsilon_{min} \right)} \tag{29}$$

subject to $x' \in N_x$ and $\hat{y}_x = \hat{y}_{x'}$.

Relative Representation Stability (RRS): RRS measures the stability relative to the model's internal latent representations $L(\cdot)$:

$$\text{RRS}(x, x', e_x, e_{x'}) = \max_{x'} \frac{\left\| \frac{e_x - e_{x'}}{e_x} \right\|_p}{\max\left(\left\| \frac{L(x) - L(x')}{L(x)} \right\|_p, \epsilon_{min} \right)} \tag{30}$$

Relative Output Stability (ROS): ROS measures the stability relative to the change in output prediction probabilities $f(\cdot)$:

$$\mathrm{ROS}(x, x', e_x, e_{x'}) = \max_{x'} \frac{\left\| \frac{e_x - e_{x'}}{e_x} \right\|_p}{\max\left(\left\| \frac{f(x) - f(x')}{f(x)} \right\|_p, \epsilon_{min} \right)} \tag{31}$$

These stability metrics provide a rigorous bound on the local sensitivity of an attribution method, ensuring that explanations remain consistent for similar instances.

3.3 Summary

In this section, we discussed existing evaluation metrics and measures in feature selection and feature attribution. We observed that the local and global perspectives which existed in the underlying process of feature selection and feature attribution methods, were also observed in the metrics design.

Beyond the Randomized Label Test discussed above, which transfers to feature selection directly as it already operates on a global, dataset-level selection, the remaining feature attribution measures would require conceptual adaptation before they could be used to evaluate feature selection algorithms. The predictive faithfulness measures, PGI and PGU, are in principle the most straightforward to adapt: since feature selection produces a single, global mask over the features rather than a per-instance explanation, one could replace the local attribution vector e_x used in these measures with the constant, dataset-wide selection mask, and perturb the selected or non-selected features of every instance accordingly. This substitution is conceptually simple and would not require any change to the perturbation procedure itself, only to the source of the importance vector.

The stability measures, RIS, RRS, and ROS, are considerably harder to adapt in practice. They are defined by comparing the explanation produced for an instance x against explanations produced for a set of perturbed neighbors $x' \in N_x$, which for a feature attribution method only costs a single additional forward pass. For a feature selection algorithm, however, there is no explanation local to a single instance; obtaining a comparable notion of stability would require re-running the entire selection procedure once per perturbed instance, which is computationally prohibitive for all but the simplest filter-based methods. An alternative would be to fix the global selection mask and only test whether the downstream prediction is stable under the perturbation, but this no longer measures anything about the selection process itself and instead collapses into a standard robustness measure of the trained model, already captured, at least indirectly, by the supervised and unsupervised performance metrics reviewed for feature selection. For stability in the presence of noise however, there are dedicated methods as discussed earlier, like the work by Nogueira [36].

Overall, this asymmetry is what motivates the direction we take in this paper: rather than attempting to adapt the point-based, instance-level measures of feature attribution, PGI, PGU, RIS, RRS, and ROS, to feature selection algorithms that do not naturally produce a per-instance explanation, we instead build on the fact that feature selection's own global measures already extend to feature attribution methods once their scores are aggregated and transformed into a top-k selection, as discussed in Section 2. This is the direction we adopt for our experimental comparison of the two families of methods: evaluating both feature selection and feature attribution methods through the same global view, using the measures established for feature selection. The one exception is the Randomized Label Test, which we retain as it is already defined at the dataset level and therefore transfers directly to feature selection without any modification.

Table 1. Characteristics of the high-dimensional benchmark datasets.

Dataset	Instances	Features
ALLAML	72	7,129
arcene	200	10,000
Carcinom	174	9,182
CLL_SUB_111	111	11,340
COIL20	1,440	1,024
colon	62	2,000
gisette	7,000	5,000
GLI_85	85	22,283
GLIOMA	50	4,434
Isolet	1,560	617
leukemia	72	7,070
lung	203	3,312

Dataset	Instances	Features
lung_discrete	73	325
madelon	2,600	500
ORL	400	1,024
orlraws10P	100	10,304
pixraw10P	100	10,000
Prostate_GE	102	5,966
SMK_CAN_187	187	19,993
TOX_171	171	5,748
warpAR10P	130	2,400
warpPIE10P	210	2,420
Yale	165	1,024

Algorithm 1 Global Feature Importance via Attribution Averaging

Input: Dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ with d features, Model f, Local Attribution Method ϕ
Output: Global Importance Vector $G \in \mathbb{R}^d$
Initialize: $G \leftarrow [0, 0, \ldots, 0]$ (zero vector of length d)
for each instance $x_i \in \mathcal{D}$ **do**
  Compute local attribution: $a_i \leftarrow \phi(f, x_i)$ ▷ a_i represents feature contributions for point x_i
  $G \leftarrow G + |a_i|$ ▷ Accumulate absolute contributions to capture magnitude
end for
$G \leftarrow \frac{1}{n} \sum_{i=1}^{n} |a_i|$ ▷ Calculate the mean across all instances
return G

4 Experimental Setup

4.1 Datasets and Characteristics

We conduct extensive experiments on a diverse set of large and high-dimensional datasets sourced from the *scikit-feature* repository [30]. These datasets span multiple domains, including genomics, proteomics, and image recognition, representing typical challenges in high-dimensional data analysis where the number of features (d) often significantly exceeds the number of instances (n). An overview of the used datasets and their characteristics is presented in Table 1.

4.2 Benchmarking and Evaluation Methodology

Our experiments are performed using the `FSEVAL` benchmarking suite [43], a specialized framework designed for the systematic and reproducible evaluation of feature selection algorithms. We evaluate both feature selection and feature attribution methods detailed in Section 2 using the global feature selection evaluation metrics surveyed in Section 3. Integrated Gradients and Sobol are excluded as they impose some limitations. Integrated Gradients only works with differentiable models, and Sobol is computationally expensive. This omission also allows a balanced comparison of 6 feature selection and 6 feature attribution methods. Let us remind the reader that we do not employ specific feature attribution evaluation metrics in this context, as feature attribution methods provide a *local perspective*, assigning different importance scores to features for each individual query point. The evaluation metrics also follow this principle, and hence, it is not feasible to employ them for the evaluation of feature selection methods providing a *global* dataset-wise feature importance value. For the experiments, we treat these feature attributions as global feature importance scores and evaluate them using standard feature selection metrics based on Alg. 1.

The evaluation protocols are defined as follows:

Supervised Evaluation: We utilize a Random Forest [9] classifier. To ensure a robust estimation of performance and mitigate variance caused by small sample sizes, we implement a 5-fold stratified cross-validation repeated 10 times.

Unsupervised Evaluation: For clustering-based assessment, we use the k-means algorithm. To account for the stochastic nature of centroid initialization, we report the average results across 10 independent runs.

Model-Agnostic Evaluation: We use Average Angle Difference (AAD) [40] to provide a model-agnostic evaluation of the methods as well.

Stability: For each of the aforementioned metrics, the stability [38] is calculated, to show the stability of a method's selection process with respect to that metric.

Scalability: We evaluate how scalable each model is by conducting an extensive scalability experiment based on both varying number of data points, and features.

4.3 Experimental Scenarios

To capture a holistic view of how selection methods behave across different scales, we conduct two distinct experimental scenarios:

(1) **General Feature Selection:** We study the overall selection process by selecting from 5% to 100% of the features, using a step size of 5%. This provides a broad view of the redundancy and information density within the datasets.

(2) **Extreme Dimensionality Reduction:** In high-dimensional contexts, practical utility often requires selecting a very small subset of features. Furthermore, high feature-to-instance ratios can sometimes mask the true performance of an algorithm. Therefore, we conduct a fine-grained experiment selecting from 0.5% to 10% of features with a step size of 0.5%. This scenario allows us to assess the effectiveness of methods in "extreme" cases where only the most informative features are retained for the final downstream task.

The aforementioned experiments are designed to evaluate and compare methods from different aspects, providing a comprehensive view of how the methods perform.

5 Comparative Analysis

For both experimental scenarios, we first provide the feature selection similarity matrices. Let M_1 and M_2 be two feature selection methods. For a given selection budget, k features to be selected, let $S_1 \subseteq F$ and $S_2 \subseteq F$ represent the sets of the top-k features selected by M_1 and M_2, respectively, where F is the total set of features. The feature selection similarity R_S at cardinality k is defined as the normalized cardinality of the intersection:

$$R_S(S_1, S_2, k) = \frac{|S_1 \cap S_2|}{k} \tag{32}$$

where $R_S \in [0, 1]$. A value of 1 indicates identical feature preferences, while a value of 0 indicates completely different selections. In our analysis, we express k as a percentage of the total feature count ($p \in \{5\%, 10\%, 15\%, 20\%, 25\%\}$ for the general feature selection scenario, and $p \in \{0.5\%, 1\%, 1.5\%, 2\%, 2.5\%\}$ for the extreme dimensionality reduction case) to observe the similarity between different feature selection and feature attribution methods.

We continue the presentation of the experimental results by providing examples of performance plots for each of the different metrics. To provide a better glance at the overall performance of the methods for each metric, we provide Critical Difference (CD) diagrams based on classic rank statistics [17] and on magnitude-aware rank statistics (MARS) [41]. Finally, we conclude the presentation of the experimental results by highlighting the scalability analysis of the selected methods.

5.1 General Feature Selection Performance

In this section, we provide a deep dive into the experimental results for the "General Feature Selection" scenario, where the selection budget spans from 5% to 100%. This wide-lens evaluation is crucial, as it captures the performance degradation or improvement as a model moves from a highly parsimonious state to full dimensionality. We utilize the 100% experiment results to illustrate these trajectories.

5.1.1 Feature Selection Similarity. The *Feature Selection Similarity* metric reflects the consensus between different algorithms. High similarity between two methods suggests they identify the same underlying signal, while low similarity indicates that they capture different facets of the informational structure of the data. We use this experiment to check if we find any interesting expected or unexpected patterns of similarity in selection between different feature selection and feature attribution algorithms.

As illustrated in Figure 1, we present the feature selection similarity heatmaps across five critical selection budgets. At the 5% threshold, the divergence between methods is most pronounced, highlighting the unique preferences of each method when forced to be highly selective. As we progress to the 30% budget, we see the increase in the similarity values, as the probability of the divergence is decreased.

An interesting observation is the high similarity of LASSO, as a feature selection method, with the Permutation, as a feature attribution method. This is the highest amount of similarity observed between any pair of methods. Methods like Tree_SHAP or Kernel_SHAP also show a higher similarity to the feature selection methods, compared to the feature

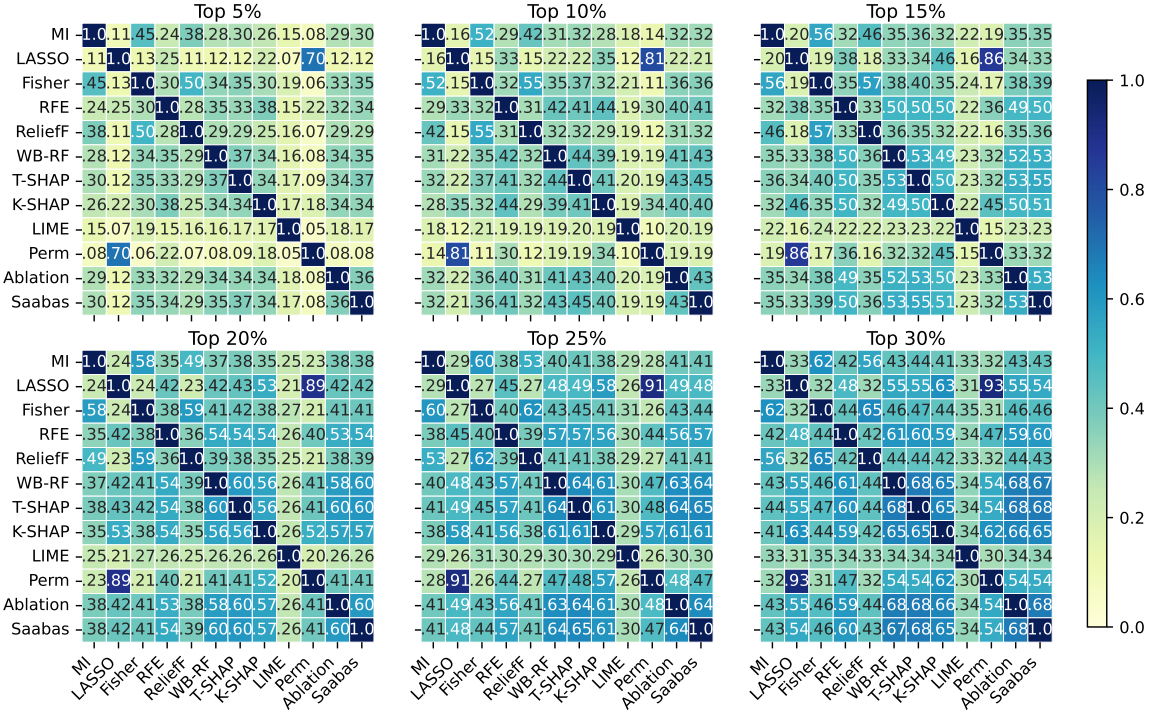


Fig. 1. Feature Selection Similarity heatmaps across different budgets. The grid illustrates the behavior of the methods compared to each other, and as the feature subset size increases.

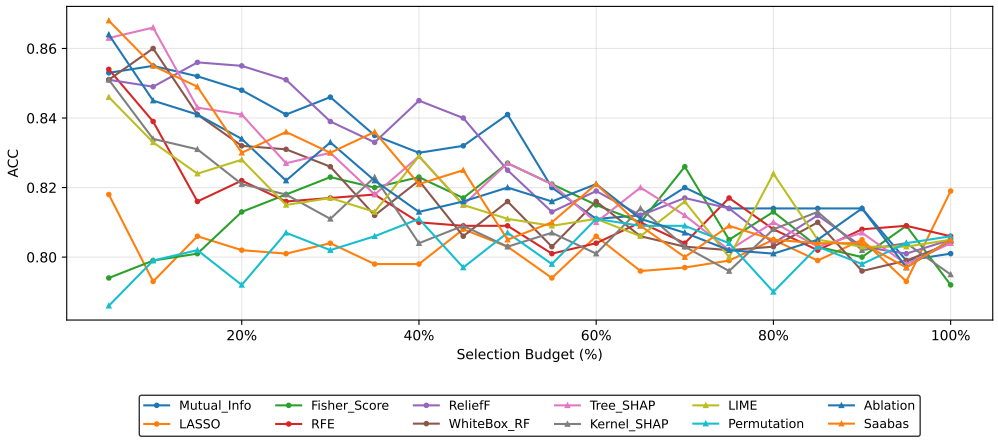


Fig. 2. Performance trajectory for Classification Accuracy (ACC) on the *arcene* dataset.

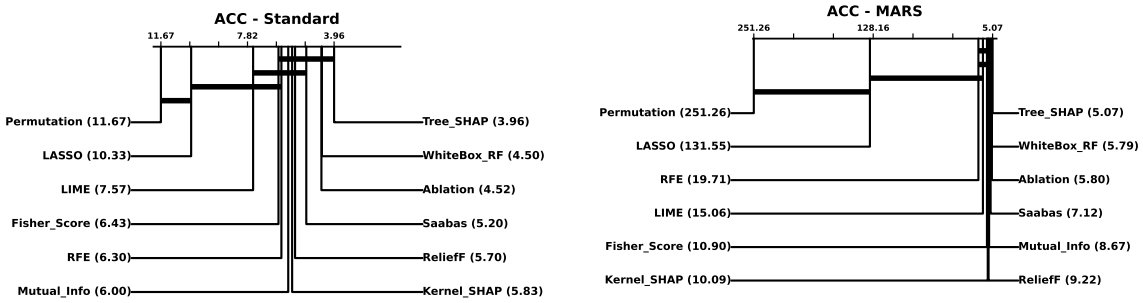


Fig. 3. Global Rank Analysis for Accuracy (ACC) across all datasets.

attribution. Wrapper-based feature selection methods show a higher-similarity to the feature attribution methods than other feature selection candidates, which is expected, as they all have Random Forest in common in their backend. Feature selection methods that do not rely on a predictive model however, show a higher similarity towards the candidates from their own family.

5.1.2 Classification Accuracy (ACC). Accuracy serves as the primary benchmark for supervised utility, alongside AUC. In high-dimensional datasets like *arcene*, which we use as an example of our dataset selection, feature selection acts as a vital regularizer. Figure 2 shows the accuracy trajectory for various methods. We often observe a "peak" early in the budget, after which the inclusion of additional, likely redundant features causes the Random Forest classifier's performance to plateau or even slightly decay due to the curse of dimensionality. The performance of the feature selection and feature attribution methods seems very much entangled and competitive.

Figure 3 provides the rank statistics analysis. By placing the Standard and MARS analyses side-by-side, we observe the overall performance of the methods based on the accuracy on all of the 23 datasets. Standard rank analysis uses the rank information only, while MARS presents the magnitude-aware perspective, by incorporating actual metric values.

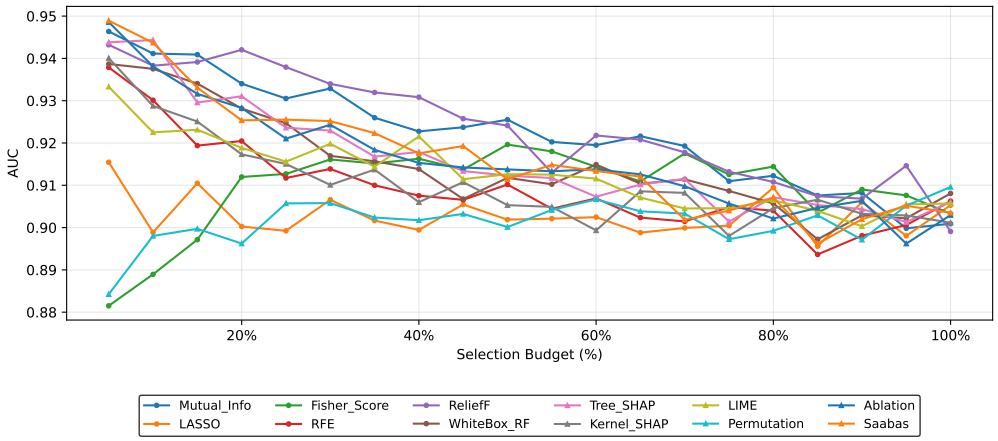


Fig. 4. Performance trajectory for Area Under Curve (AUC) on the *arcene* dataset.

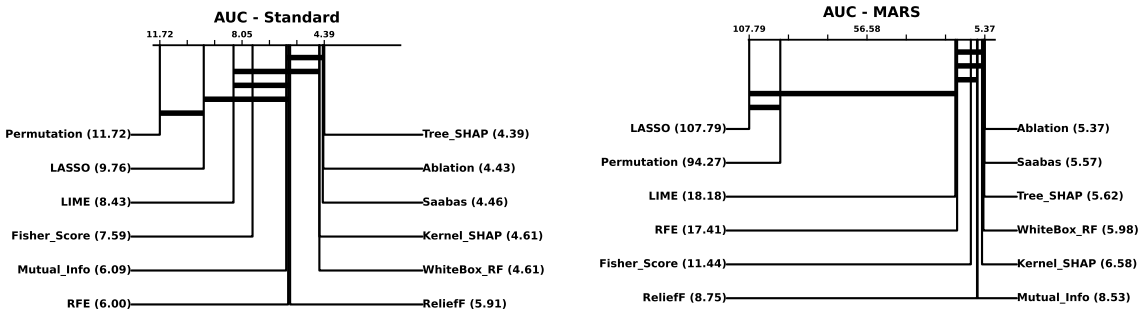


Fig. 5. Global Rank Analysis for Area Under Curve (AUC) across all datasets.

In both rank statistics, feature attribution methods form most of the top clusters, and the best feature selection method is WhiteBox_RF, which is a wrapper-based method based on Random Forest. The interesting observation is, that despite being designed for the provision of local feature attributions as a form of explanation, the global perspective derived from these methods is still, on average, more effective than feature selection techniques.

5.1.3 Area Under Curve (AUC). The Area Under the ROC Curve (AUC) offers a threshold-independent measure of discriminative power. As shown in the performance line plot in Figure 4, the same trend as ACC is observed in AUC as well. These experiments show that, at least for the supervised tasks, there are a lot of redundancies in the features, as the supervised metrics are degrading as more features are selected.

In the rank analysis (Figure 5), we observe a quite similar figure to the ACC, with the exception of a few changes in positions. One interesting observation to take away from the supervised experiments is the performance of LIME. As the representative for the surrogate feature attribution methods, LIME performs very poorly. MARS reflects that by considering the metric values in the rank analysis, we can see that it is close to the *winners* cluster. However, it is still the worst performing feature attribution method, if we disregard Permutation. This might be due to the use of a

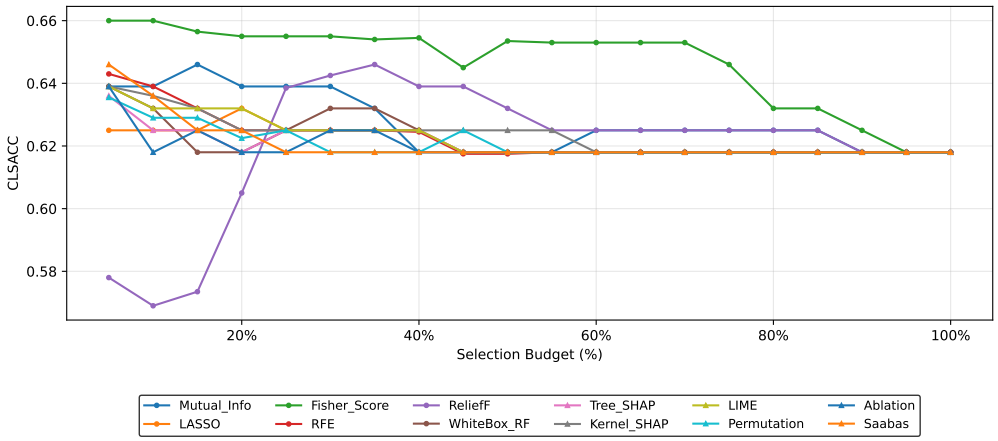


Fig. 6. Performance trajectory for Clustering Accuracy (CLSACC) on the *arcene* dataset.

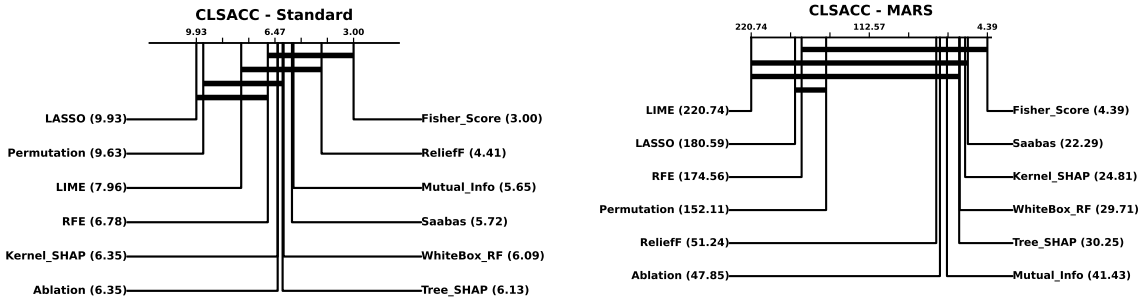


Fig. 7. Global Rank Analysis for Clustering Accuracy (CLSACC) across all datasets.

surrogate model and the amplified notion of locality in LIME's feature attributions, which renders it less effective for global feature importance values.

5.1.4 Clustering Accuracy (CLSACC). Clustering Accuracy (CLSACC) provides an unsupervised perspective by measuring how well the selected features reconstruct the ground-truth groupings of the data. Figure 6 reveals the unsupervised performance of the evaluated methods. Interestingly, it seems that feature selection methods perform better based on Clustering accuracy, with the Fisher showing clear dominance based on the dataset we used as an example, *arcene*.

The rank analysis in Figure 7 reflects that based on standard rank analysis, feature selection methods show a clear advantage over feature attribution methods. But MARS analysis shows that by incorporating metric values, the ranking changes. Fisher remains the best, but it is followed by Saabas and Kernel SHAP. This highlights that feature attribution methods, despite the fact that they might be ranked poorly in different cases, they still show competitive measures, and when they show a better rank, their margin of superiority is higher, compared to a method such as ReliefF. This demonstrates the better robustness and confidence of feature attribution methods in performing feature selection and dimensionality reduction for unsupervised tasks.

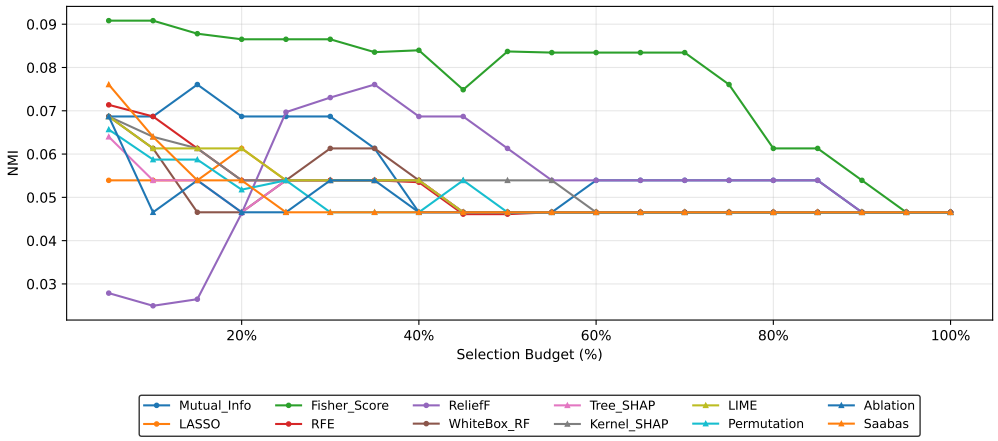


Fig. 8. Performance trajectory for Normalized Mutual Information (NMI) on the *arcene* dataset.

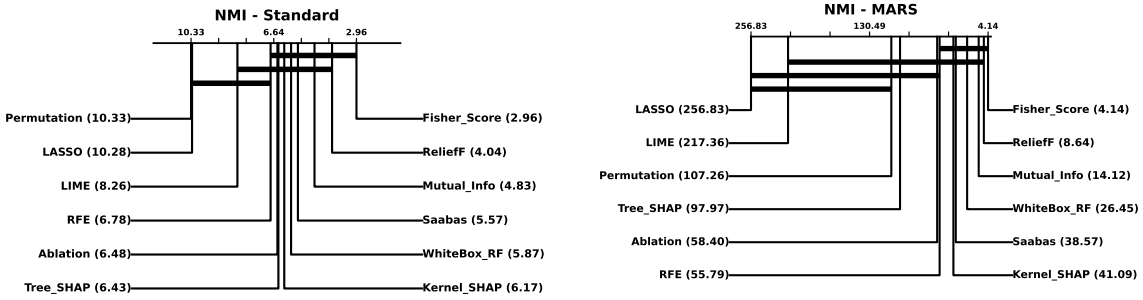


Fig. 9. Global Rank Analysis for Normalized Mutual Information (NMI) across all datasets.

5.1.5 Normalized Mutual Information (NMI). Normalized Mutual Information (NMI) serves as a robust, information-theoretic measure of clustering quality. Unlike CLSACC, NMI is less sensitive to cluster-label permutations. Figure 8 depicts the NMI performance, which, on *arcene*, is still quite similar to the figures observed for CLSACC.

The global rank analysis in Figure 9 provides a definitive view of which methods are most capable of preserving the informational "essence" of the datasets. In the case of NMI, the agreement between standard rank analysis and MARS is higher, reflecting the same three methods as top. This shows that, based on the ability to cluster points in the correct cluster as denoted by ground-truth labels, from an information-theoretic point of view feature selection methods show a clear advantage over feature attribution techniques.

5.1.6 Average Angle Difference (AAD). Average Angle Difference (AAD) is a model-agnostic metric which is employed to measure the consistency of the importance scores assigned to the features across the selection budget in preserving the geometry. Figure 10 shows the performance plot. AAD on the *arcene* dataset shows permutation as the best method. It seems, at first-glance, that this observation is the opposite of what we observed earlier. But there is a clear explanation [40]: AAD favors the methods successfully preserving the geometrical properties of the data. Recalling that, in high-dimensional spaces, distances are loosing contrast and concepts such as Nearest Neighbors are obscured, we

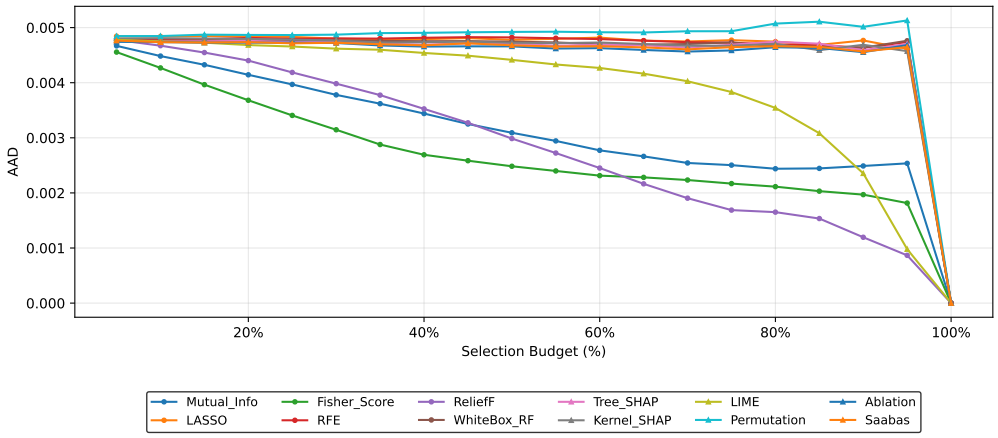


Fig. 10. Performance trajectory for Average Absolute Difference (AAD) on the *arcene* dataset.

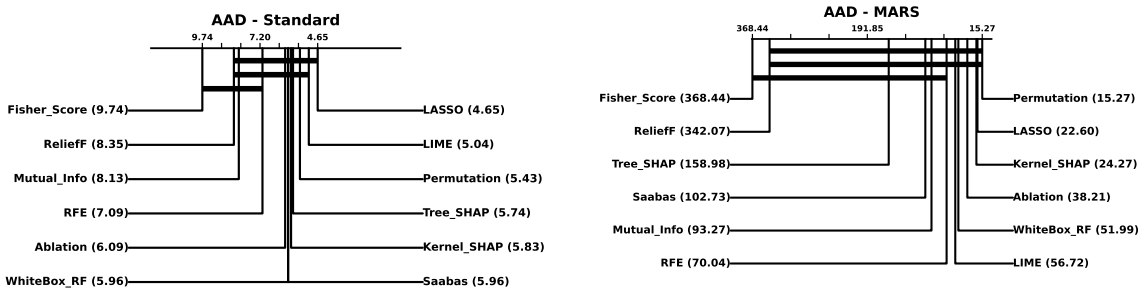


Fig. 11. Global Rank Analysis for Average Absolute Difference (AAD) across all datasets.

need to break down the geometrical properties of the data in order to get better results for supervised and unsupervised tasks. Therefore, AAD is likely to show values very different from the other metrics.

This phenomenon is also reflected in the global ranks, as shown in Figure 11. The critical difference diagrams seem to be reversed, compared to the previously presented metrics.

5.2 Extreme Dimensionality Reduction

In this section, we provide a deep dive into the experimental results for the "Extreme Dimensionality Reduction" scenario, where the selection budget spans from 0.5% to 10%. These sets of experiments can be seen as more informative and important, as they follow a more realistic scenario in which we aim to reduce the dimensionality dramatically. This challenges and evaluates the true strength of feature selection and attribution methods in selecting the best possible set of features, and gives us a closer look at the effectiveness of the features they consider as the most important.

5.2.1 Feature Selection Similarity. In Figure 12, we present the feature selection similarity heatmaps across five critical selection budgets. At the 0.5% threshold, the divergence between methods is most pronounced, highlighting the unique

preferences of each ranker when forced to be highly selective. As we progress to the 3% budget, we see the increase in the similarity values, as the probability of the divergence is decreased.

An interesting observation, just like in the general selection scenario, is the high similarity of LASSO, as a feature selection method, with the Permutation, as a feature attribution method. This is the highest amount of similarity observed between any pair of methods. Methods like Tree_SHAP or Kernel_SHAP also show a higher similarity to the feature selection methods, compared to feature attribution. Wrapper-based feature selection methods show a higher-similarity to the feature attribution methods than other feature selection candidates, which is expected, as they all have Random Forest in common in their backend. Feature selection methods that do not rely on a predictive model however, show a higher similarity towards the candidates from their own family. The similarity between the same group of methods is higher in this scenario, compared to the general feature selection.

5.2.2 Classification Accuracy (ACC). Figure 13 shows the accuracy trajectory for various methods. Unlike the general scenario, here we mostly see an upward trend, with the feature attribution methods performing generally well.

Figure 14 provides the rank statistics analysis based on the ACC. In both rank statistics, feature attribution methods form most of the top clusters, but the top two methods are from the feature selection family. They however, only win by a very low margin, especially according to MARS. Again, the only feature attribution method which appears in the *losers* cluster, is LIME, which is unable to provide effective global importance values due to its extremely local perspective.

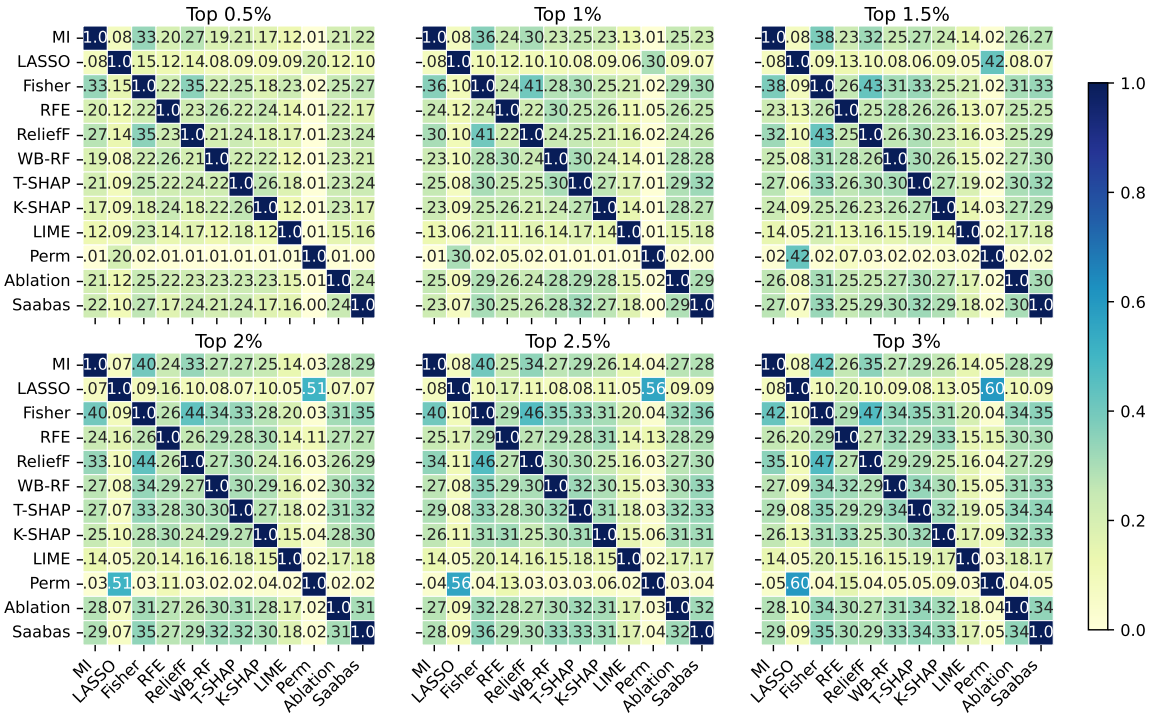


Fig. 12. Feature Selection Similarity heatmaps across different budgets. The grid illustrates the behavior of the methods compared to each other, and as the feature subset size increases.

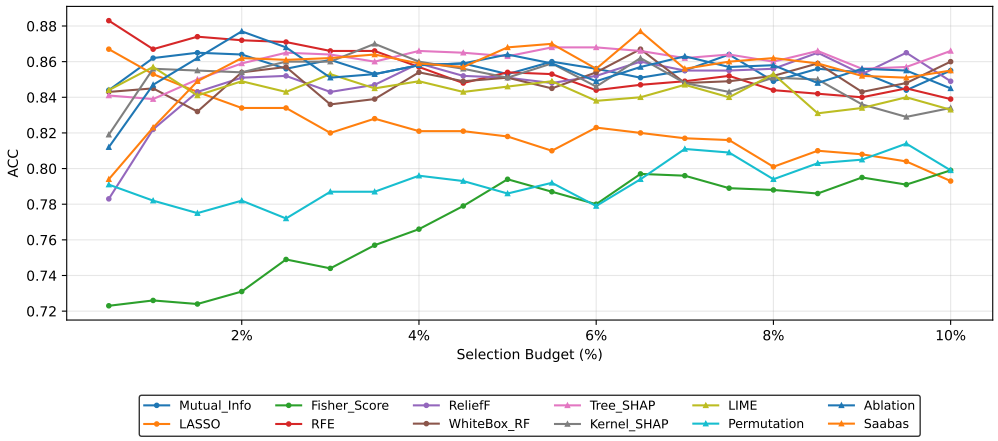


Fig. 13. Performance trajectory for Classification Accuracy (ACC) on the *arcene* dataset.

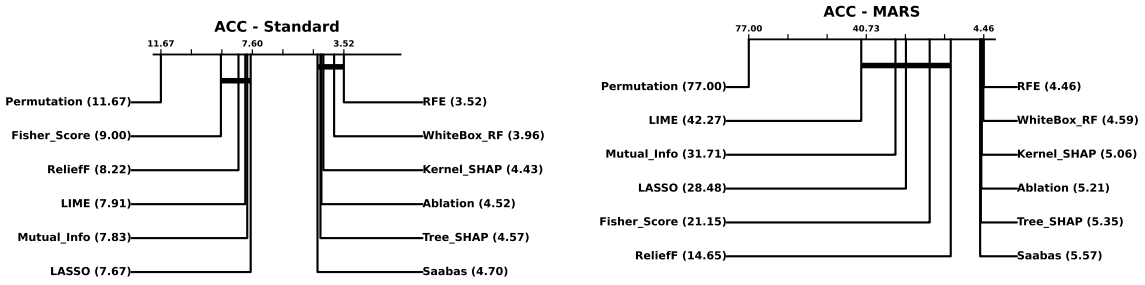


Fig. 14. Global Rank Analysis for Accuracy (ACC) across all datasets.

5.2.3 Area Under Curve (AUC). Figure 15 illustrates the accuracy trajectories for the different methods. In contrast to the general scenario, the curves mostly exhibit an upward trend, and feature attribution approaches tend to perform well, and better overall. Figure 16 presents the AUC-based rank statistics and analysis. Again, just like ACC, feature attribution methods dominate the top clusters in both ranking summaries, although the two highest-ranked methods belong to the feature selection family. Their advantage, however, is small, particularly according to MARS. As before, LIME is the only feature attribution method appearing in the losers cluster, reflecting its limited ability to produce reliable global importance values due to its strongly local perspective.

5.2.4 Clustering Accuracy (CLSACC). Figure 17 reveals the unsupervised utility of the evaluated methods. Interestingly, it seems that feature selection methods perform better based on Clustering accuracy, with the Fisher Score showing clear dominance based on the dataset we used as an example, *arcene*.

The rank analysis in Figure 18 interestingly shows a strong difference between standard and magnitude-aware rank analyses, with the best method changing and some of the methods such as ReliefF going through dramatic changes. The overall picture however, reflects that again, feature attribution methods other than LIME are actually very good in terms of CLSACC, forming the majority of the top cluster.

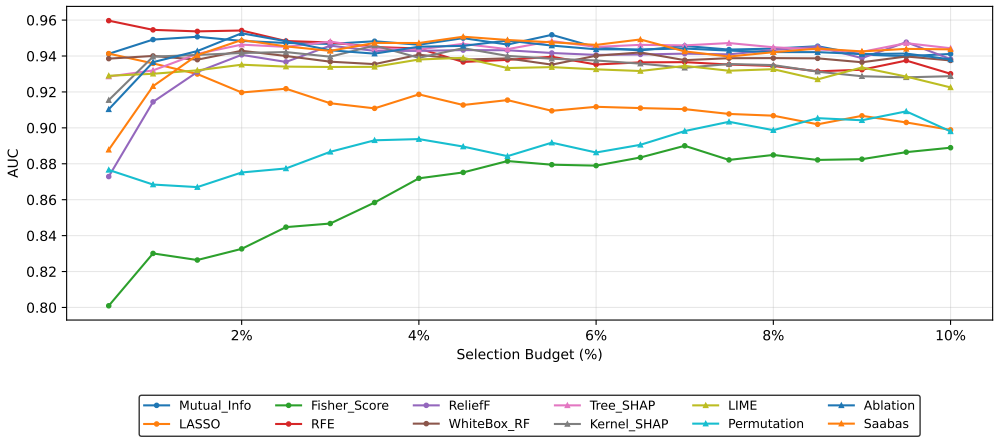


Fig. 15. Performance trajectory for Area Under Curve (AUC) on the *arcene* dataset.

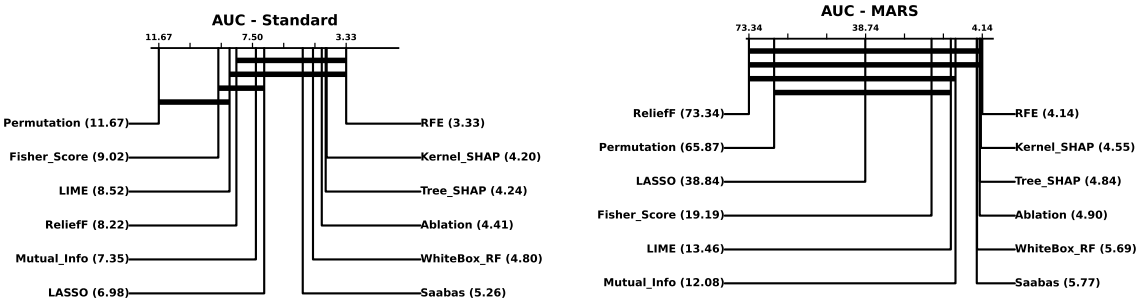


Fig. 16. Global Rank Analysis for Area Under Curve (AUC) across all datasets.

5.2.5 Normalized Mutual Information (NMI). Figure 19 depicts the NMI performance. It is, on *arcene*, still quite similar to the figures observed for CLSACC.

The global rank analysis in Figure 20 provides a definitive view of which methods are most capable of preserving the informational "essence" of the datasets. In the case of NMI, the agreement between standard rank analysis and MARS is lower, compared to the general scenario. The top two methods alternate, and the rest also change. However, according to both methods, there is an agreement between feature selection and attribution in the top cluster.

5.2.6 Average Angle Difference (AAD). Figure 21 shows the performance plot. AAD on the *arcene* dataset shows Fisher Score as the most influential method for preserving the geometrical properties of the data. However, as discussed also for the general case, the trend seems to be very different, compared to the other metrics. This phenomenon is also reflected in the global ranks, as shown in Figure 22. The critical difference diagrams seem to be reversed, compared to the previously presented metrics, with LIME appearing as the best based on both rank statistics.

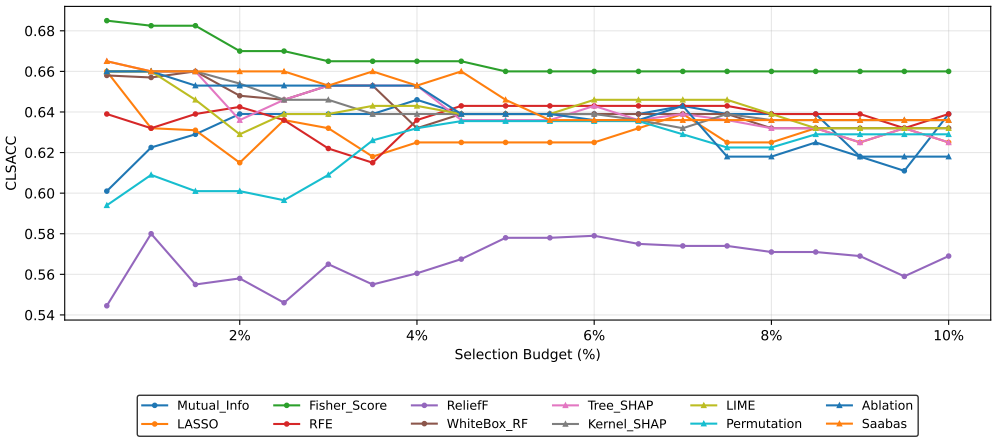


Fig. 17. Performance trajectory for Cluster Accuracy (CLSACC) on the *arcene* dataset.

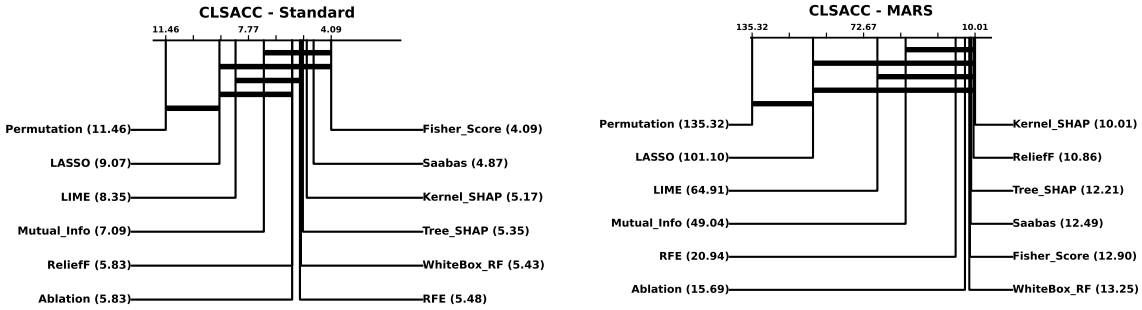


Fig. 18. Global Rank Analysis for Cluster Accuracy (CLSACC) across all datasets.

5.3 Time and Scalability Analysis

In this section, we assess computational overhead of the evaluated methods across varying scales of features (d) and instances (n). We distinguish between **Feature Selection** (the first six methods) and **Feature Attribution** (the subsequent six methods).

The empirical results documented in Figure 23 demonstrate a clear hierarchy in performance. Feature selection filters such as *Fisher Score* and *Mutual Information* are highly efficient, scaling linearly and remaining computationally negligible. However, *Recursive Feature Elimination (RFE)* and most **Feature Attribution** methods (like *Kernel SHAP* and *LIME*) exhibit substantial growth in runtime. For attribution methods, the primary cost arises from the requirement to compute local importance for every instance to derive a global score, a process that becomes prohibitively slow on high-dimensional datasets without parallelization.

6 Discussion

In this section, we provide insights based on our empirical findings. By evaluating six from the traditional Feature Selection family and six from the modern Feature Attribution family across 23 high-dimensional datasets, we can now

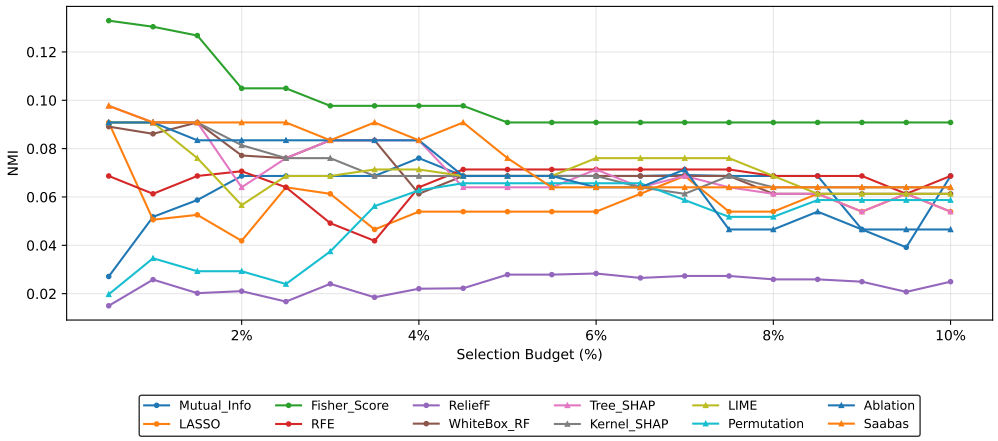


Fig. 19. Performance trajectory for Normalized Mutual Information (NMI) on the *arcene* dataset.

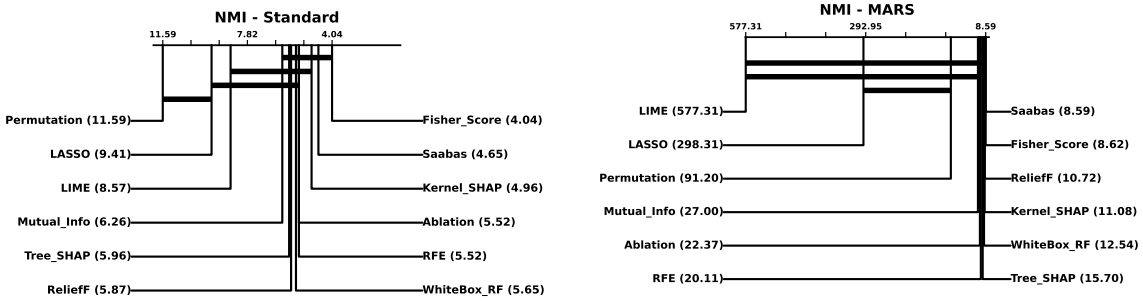


Fig. 20. Global Rank Analysis for Normalized Mutual Information (NMI) across all datasets.

address the fundamental tension between the principle of parsimony and the demand for accountability. The following analysis aims to answer Research Question 4.

6.1 The Supervised Superiority of Accountability

One of the most consistent patterns observed in our results is the dominance of feature attribution methods in supervised tasks (ACC and AUC). Theoretically, feature selection filters like *Fisher Score* or *Mutual Information* operate on univariate or simple multivariate statistical properties. In contrast, feature attribution methods like *Tree SHAP*, *Saabas*, and *Ablation* are "interrogators" of a pre-trained, complex model (in our case, Random Forest).

Our results suggest that the "What did the model actually use?"-perspective that feature attribution methods provide, is actually a more effective proxy for classification utility than the "What do I need to know?"-perspective of feature selection methods. Because feature attribution methods capture the high-order interactions and non-linear boundaries learned by the Random Forest (or any other complex or black-box predictive model), the features they prioritize are inherently optimized for the downstream task. This confirms that, while feature selection seeks an objective truth within the data, feature attribution seeks a functional truth within the model. The fact that *WhiteBox_RF* (a wrapper-based

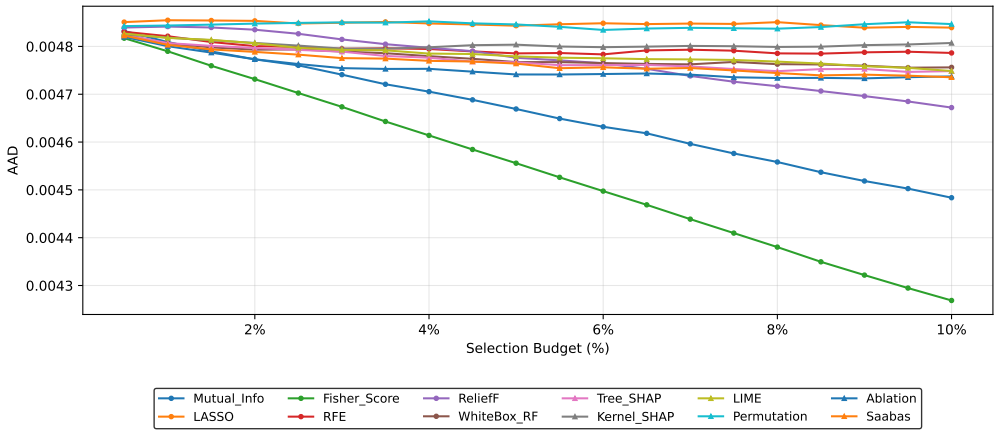


Fig. 21. Performance trajectory for Average Absolute Difference (AAD) on the *arcene* dataset.

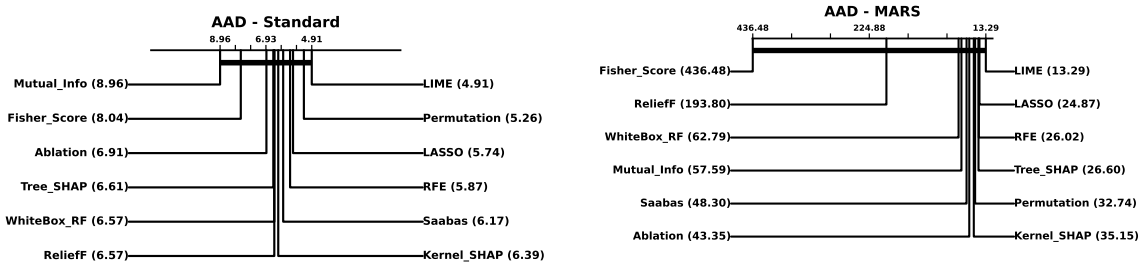


Fig. 22. Global Rank Analysis for Average Absolute Difference (AAD) across all datasets.

feature selection method) was the only feature selection method to consistently compete in the top cluster reinforces this: when selection is tied to the model's internal logic, performance improves.

6.2 The Unsupervised Resilience of Parsimony

The narrative shifts dramatically when moving to unsupervised tasks (CLSACC and NMI). Here, traditional feature selection methods, particularly *Fisher Score* and *ReliefF*, demonstrated remarkable resilience and often outperformed their feature attribution counterparts in standard rank analysis.

This divergence reveals a fundamental "model-bias" inherent in feature attribution. A feature attribution method is only as good as the model it explains. If a predictive model (e.g., Random Forest) achieves high accuracy by focusing on a few highly discriminative features while ignoring the broader manifold structure of the data, the resulting attribution will be narrow. For unsupervised clustering, which requires the preservation of the data's global informational "essence," the model-agnostic, statistical approach of feature selection filters proves superior. They are not blinded by a specific classification objective and are thus more capable of preserving the natural groupings of the data. This provides a clear answer to the applicability of criteria across groups: feature selection is better for data exploration, while feature attribution is better for task-specific optimization.

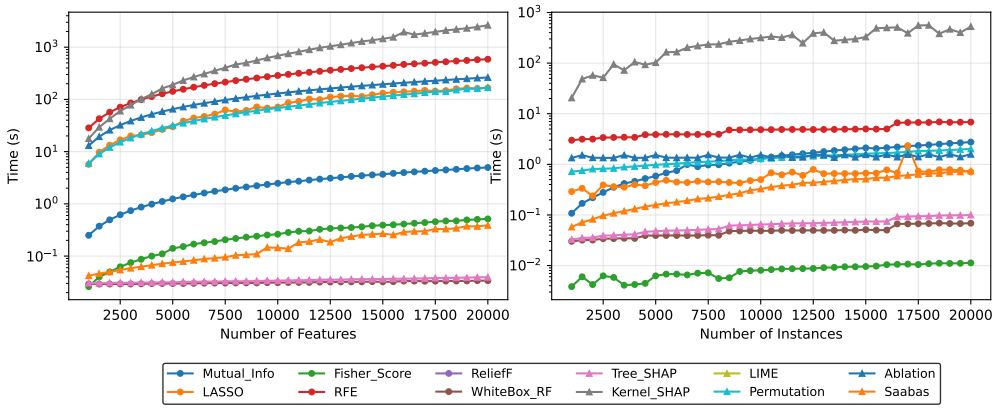


Fig. 23. Computational scalability analysis showing runtime (log-scale) as a function of feature and instance counts. The results highlight the efficiency of statistical filters compared to the high cost of iterative wrappers and attribution globalization.

6.3 Geometrical Preservation vs. Task Utility

The *Average Angle Difference (AAD)* results offer perhaps the most intriguing insight into the *curse of dimensionality*. We observed that methods which performed best in supervised tasks often performed poorest in preserving the raw geometry of the data.

This phenomenon suggests that in high-dimensional spaces, the "objective truth" of the original feature space is often a hindrance. To achieve high predictive accuracy, a model must effectively "break" the original geometry to find a lower-dimensional manifold where classes are separable. Therefore, feature attribution methods, which follow the model's lead, naturally prioritize features that transform the space rather than those that preserve its raw state. This highlights a philosophical conflict: is the "correct" feature one that reflects the raw data (feature selection) or one that enables a decision (feature attribution)? Our results favor the latter for utility and the former for structural integrity.

6.4 Convergence: Where Two Sides of the Coin Meet

Despite their different starting points, our similarity heatmaps (Figures 4 and 9) show significant "epistemological agreement" between certain pairs. The high similarity between *LASSO* (FS) and *Permutation* (FA) is particularly noteworthy. *LASSO* uses a penalty term to force simplicity (parsimony), while *Permutation* measures the loss of performance when a feature is removed (accountability).

Their convergence suggests that for many high-dimensional problems, the features that are most necessary for a sparse model are the same ones that are most impactful when removed from a complex one. This overlap provides a bridge for researchers: if a practitioner needs the speed of a feature selection filter but the rationale of a feature attribution method, choosing a method like *LASSO* might provide a mathematically grounded middle ground.

6.5 Practical Implications: The Hybrid Pipeline

Based on the scalability analysis in Section 5.3, we propose that the modern high-dimensional data mining and machine learning pipeline should not choose between feature selection and attribution methods, but rather utilize them sequentially. The univariate efficiency of feature selection filters should be used to eliminate the "noise" (parsimony),

reducing the features from tens of thousands to hundreds. Subsequently, the high-fidelity interrogation process of attribution methods should be used to select the final subset that the model actually utilizes (accountability). This hybrid approach respects the computational constraints of Big Data while leveraging the explanatory power of modern XAI.

7 Summary and Conclusion

In this study, we have navigated the complex and often overlapping landscapes of feature selection and feature attribution, two methodologies that serve as the bedrock of high-dimensional data analysis. Traditionally, these fields have operated in relative isolation: feature selection, born from the statistical principle of parsimony, focuses on the pursuit of an objective, minimal subset of data that retains predictive or descriptive power; feature attribution, a child of the modern Explainable AI movement, focuses on the pursuit of accountability, interrogating complex models to reveal the internal logic behind their decisions. By placing twelve fundamental feature selection and feature attribution methods ranging from univariate filters like Fisher Score to advanced model-agnostic explainers like Kernel SHAP on a common analytical scale, we have bridged a significant gap in the literature. Our evaluation across 23 diverse datasets and two distinct selection scenarios (General and Extreme) has allowed us to move beyond philosophical definitions, into a rigorous, empirical comparison of how these *two sides of the same analytical coin* actually perform in practice.

The findings of this paper reveal a nuanced hierarchy of utility. In supervised tasks, our results provide a clear mandate for the feature attribution family. Methods such as *Tree SHAP*, *Saabas*, and *Ablation* consistently dominated the rank statistics for Accuracy and AUC, suggesting that the functional truth learned by a high-performing model (like Random Forest) is a more effective guide for feature importance than raw statistical filters. This marks a critical shift: if the goal is predictive performance, asking *"What did the model actually use?"* is often more informative than asking *"What is objectively relevant?"*. However, this dominance is inverted in unsupervised contexts. In clustering tasks (CLSACC and NMI), the statistical filters of the feature selection family demonstrated superior robustness. Because feature selection filters are not blinded by a specific classification objective, they remain more faithful to the data's natural manifold, preserving informational structures that task-specific attribution methods might discard as noise.

Furthermore, our exploration of the *Average Angle Difference (AAD)* and scalability benchmarks exposed the physical and geometrical limits of these methods. We observed a *performance-geometry paradox* where the features prioritized by high-utility methods often actively distorted the original data geometry to achieve class separability. On the computational front, our scalability analysis confirmed a significant barrier to the globalization of feature attribution; while feature selection filters scale linearly and effortlessly with dimensionality, the globalization cost of local attributions creates a computational bottleneck that necessitates a tiered approach to Big Data. Ultimately, this survey proves that while feature selection and feature attribution methods sometimes and partially converge on the same key features (as seen in the high similarity between LASSO and Permutation), they offer distinct advantages: feature selection provides a scalable, preservative foundation, while feature attribution provides a high-fidelity, performance-oriented interrogation.

Despite the comprehensive nature of this study, several implications remain that must be carefully declared and highlighted to pave the way for future inquiry. First, our empirical evaluation relied heavily on Random Forest as the primary backend for both wrapper-based selection and tree-specific attribution. While this ensured a fair comparison between families, the behavior of these methods when paired with different architectures—such as Gradient Boosted Trees or Deep Neural Networks, might be different. The aim of this comparative study was to provide insights and shed light on the similarities and differences of the feature selection and feature attribution methods, not to stress-test their performances in different scenarios. In addition, our globalization of local attribution methods was performed through

simple averaging (mean absolute attribution); alternative aggregation strategies might yield different global rankings, which might be worth exploring.

We suggest that future work, especially in the fields concerned with very high-dimensional data, should focus on the development of hybrid selection frameworks that formally integrate the efficiency of feature selection methods with the fidelity and the effectiveness of the feature attribution techniques into a single, automated pipeline. There is also a significant opportunity to explore Multi-Objective Feature Selection, where methods are optimized simultaneously for supervised accuracy, unsupervised cluster preservation, and geometrical consistency. Additionally, extending this comparative framework to regression tasks and time-series data would broaden the applicability of our findings. By continuing to break down the silos between statistical parsimony and algorithmic accountability, the field can move toward a more transparent and effective era of high-dimensional machine learning.

References

[1] Kumar Abhishek and Deeksha Kamath. 2022. Attribution-based XAI methods in computer vision: A review. *arXiv preprint arXiv:2211.14736* (2022).

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 31. 9505–9515.

[4] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 15784–15799.

[5] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).

[6] Roberto Battiti. 1994. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Trans. Neural Networks* 5, 4 (1994), 537–550. doi:10.1109/72.298224

[7] Richard Bellman. 1966. *Dynamic programming*. Science.

[8] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. 1987. Occam's razor. *Information processing letters* 24, 6 (1987).

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[10] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees*. CRC Press, Boca Raton, FL.

[11] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & electrical engineering* 40, 1 (2014), 16–28.

[12] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York.

[13] Thomas M Cover and Joy A Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons.

[14] David R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958).

[15] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: a survey. *arXiv preprint arXiv:2105.11266* (2021).

[16] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[17] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *JMLR* 7 (2006), 1–30. doi:10.5555/1248547.1248548

[18] Pradip Dhal and Chandrashekhar Azad. 2022. A comprehensive survey on feature selection in the various fields of machine learning. *Applied intelligence* 52, 4 (2022), 4543–4581.

[19] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance Across an Entire Class of Model. *Journal of Machine Learning Research* 20, 177 (2019), 1–81.

[20] Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.

[21] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[22] Quanquan Gu, Zhenhui Li, and Jiawei Han. 2011. Generalized Fisher Score for Feature Selection. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*. 266–273.

[23] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003).

[24] Isabelle Guyon, Jason Weston, Stephen D. Barnhill, and Vladimir Vapnik. 2002. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* 46, 1-3 (2002), 389–422. doi:10.1023/A:1012487302797

[25] Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97, 1-2 (1997), 273–324.

[26] Igor Kononenko. 1994. Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning*. 171–182.

[27] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data* 3, 1 (2009), 1:1–1:58.

[28] Ludmila I. Kuncheva. 2007. A stability index for feature selection. In *IASTED International Conference on Artificial Intelligence and Applications*.

[29] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–45.

[30] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2018. Feature Selection: A Data Perspective. *ACM Comput. Surv.* 50, 6 (2018), 94:1–94:45. doi:10.1145/3136625

[31] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).

[32] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[33] Jianyu Miao and Lingfeng Niu. 2016. A survey on feature selection. *Procedia computer science* 91 (2016), 919–926.

[34] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[35] Werner Mostert, Katherine M. Malan, and Andries P. Engelbrecht. 2021. A Feature Selection Algorithm Performance Metric for Comparative Analysis. *Algorithms* 14, 3 (2021), 100.

[36] Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. 2017. On the stability of feature selection algorithms. *J. Mach. Learn. Res.* (Jan. 2017).

[37] Christos H. Papadimitriou and Kenneth Steiglitz. 1982. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall.

[38] Muhammad Rajabinasab, Anton Danholt Lautrup, Tobias Hyrup, and Arthur Zimek. 2024. A Dynamic Evaluation Metric for Feature Selection. In *SISAP (Lecture Notes in Computer Science, Vol. 15268)*. Springer, 65–72.

[39] Muhammad Rajabinasab, Anton D. Lautrup, Peter Schneider-Kamp, and Arthur Zimek. 2026. Towards Semi-supervised Subspace Learning for Outlier Detection in Big Data. In *Similarity Search and Applications*. 330–344.

[40] Muhammad Rajabinasab, Anton Danholt Lautrup, and Arthur Zimek. 2025. *Metrics for Inter-Dataset Similarity with Example Applications in Synthetic Data and Feature Selection Evaluation*. SIAM, Philadelphia, PA, USA, 527–537. doi:10.1137/1.9781611978520.57

[41] Muhammad Rajabinasab, Afsaneh M. Nejad, and Arthur Zimek. 2026. MARS: Magnitude-Aware Rank Statistics. *CoRR* abs/2605.23563 (2026).

[42] Muhammad Rajabinasab and Arthur Zimek. 2026. An Empirical Study of Feature Selection Granularity. *CoRR* abs/2607.24145 (2026).

[43] Muhammad Rajabinasab and Arthur Zimek. 2026. FSEVAL: Feature Selection Evaluation Toolbox and Dashboard. *CoRR* abs/2604.18227 (2026).

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

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *SIGKDD*.

[46] Ando Saabas. 2014. Interpreting random forests.

[47] Lloyd S Shapley et al. 1953. A value for n-person games. (1953).

[48] Ilya M Sobol. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation* 55, 1-3 (2001), 271–280.

[49] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41, 3 (2014), 647–665.

[50] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*.

[51] J. Tang and H. Alelyani, S.and Liu. 2014. Feature selection for classification: A review. *Data classification: Algorithms and applications* (2014), 37.

[52] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 58, 1 (1996), 267–288.

[53] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems* 32, 11 (2020), 4793–4813.

[54] Corne Van Zyl, Xianming Ye, and Raj Naidoo. 2024. Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy* 353 (2024), 122079.

[55] Bing Xue, Mengjie Zhang, Will N Browne, and Xin Yao. 2015. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on evolutionary computation* 20, 4 (2015), 606–626.

[56] Jan Zacharias, Moritz von Zahn, Johannes Chen, and Oliver Hinz. 2022. Designing a feature selection method based on explainable artificial intelligence. *Electronic Markets* 32, 4 (2022), 2159–2184.

[57] Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher. 2022. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEe Access* 10 (2022), 93104–93139.

[58] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.* 5, 5 (2012), 363–387.