*applied sciences*

MDPI

*Article*

# G-XCI: A Conditional Multi-View Audit of Stability and Cross-View Rank Agreement in Tabular Feature Attributions

**Chien-Hung Lai [1,*], Wen-Chun Yang [2], Yi Lin [3] and Yuh-Shyan Hwang [1]**

[1] Department of Electronic Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; yshwang@ntut.edu.tw
[2] Edrington Taiwan Ltd., 9F, No. 9 Songgao Road, Xinyi District, Taipei 11073, Taiwan; wendy.yang@edrington.com
[3] Department of Business Administration, Takming University of Science and Technology, Taipei 11451, Taiwan; linyi@takming.edu.tw
* Correspondence: ee12245@ntut.edu.tw

**Abstract**

Post hoc feature attributions are often interpreted from a single run, although their reproducibility and agreement with other explanatory views can depend on resampling, model family, representation, audit lens, output orientation, and inspection scope. We evaluate G-XCI as a conditional three-readout audit protocol. LAS measures reproducibility of an aggregate ranking derived from held-out Monte-Carlo Shapley attributions; GRS summarizes cross-seed variance of a specified global score profile; SCC measures Kendall tau-b agreement between the aggregate attribution ranking and a chosen global ranking. Across three tabular datasets, three model families, five stratified repeats, four global audit lenses, three inspection scopes, GRA ablations, and representation variants, Adult Income changed SCC sign with K, while audit-lens and representation changes materially altered SCC. Targeted supplementary audits showed that complementing the WDBC binary output reversed the mean GRA-based SCC sign in all nine model-scale combinations while leaving absolute local Shapley rankings unchanged, whereas forcing Standard GRA, absolute Spearman-to-output, and output permutation onto the same 128 held-out cases preserved all 27 original mean-SCC signs. A short lemma shows that robust-z followed by min-max equals direct min-max for nonconstant sequences. In the tested tabular binary-classification settings, G-XCI supports conditional auditing of aggregate-ranking reproducibility and cross-view rank agreement under a declared audit contract; it does not establish universal explanation correctness.

**Keywords:** explainable artificial intelligence; explanation reliability; SHAP; grey relational analysis; rank consistency; representation sensitivity; tabular machine learning

Check for updates

Academic Editor: Luis Javier García Villalba

Received: 28 August 2026
Revised: 21 September 2026
Accepted: 23 September 2026
Published: 30 September 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Explainable artificial intelligence (XAI) has become an important layer of model inspection in domains where predictions must be reviewed, justified, or communicated to human decision-makers. Post hoc feature-attribution methods are especially attractive for tabular models because they translate a prediction into feature-level contributions that can be ranked and inspected. SHAP provides a Shapley value-based additive attribution framework, while LIME and Integrated Gradients represent alternative local explanation families with different assumptions and computational mechanisms [1–3]. Despite their

practical usefulness, the existence of an explanation does not by itself establish that the explanation is reliable.

A central difficulty is that explanation quality is multi-dimensional. A single attribution plot may appear plausible while remaining sensitive to the trained model, perturbation process, sampling budget, or explanation operator. Sanity-check and benchmarking studies have therefore argued for explicit evaluation of explanation behavior rather than visual acceptance of a single run [4–6]. Work on adversarial fragility, uncertainty-aware explanations, stabilized local explanations, multi-metric evaluation, and recent Kernel-SHAP stability analysis reinforces the same point: reproducibility, robustness, faithfulness, and agreement are related but non-identical properties [7–11]. This distinction is especially important when practitioners inspect feature orderings rather than only raw attribution magnitudes.

Feature dependence and representation introduce another layer of uncertainty. Shapley-based explanations are affected by the way feature dependence is modeled, and naïve independence assumptions can change the meaning of the resulting attributions [12,13]. In tabular pipelines, the explanatory object is also conditioned on preprocessing choices such as one-hot encoding, reference-category construction, collinearity handling, and numerical scaling. Empirical studies have shown that correlated predictors, categorical/numerical composition, and preprocessing choices can materially affect local explanation behavior [14–16]. Consequently, an explanation should not be treated as a property of the trained predictor alone; it is produced by a larger analytical contract that includes the model, the representation, and the explainer.

A further issue arises when local feature attributions are compared with a second explanatory view. Different explanation paradigms need not produce the same feature ordering, and disagreement does not by itself identify which view is correct [17]. This is an important limitation of any attempt to reduce explanation reliability to a single scalar notion of correctness. An aggregate ranking derived from local attributions may be highly reproducible across repeated executions yet disagree with a stable population-level relationship; conversely, a global profile may be stable while the aggregate attribution ranking remains variable. These are operationally different failure modes and should be measured separately.

G-XCI addresses this narrower engineering problem by treating explanation reliability as a conditional multi-view audit rather than a truth-estimation task. The framework combines three complementary readouts. Local Attribution Stability (LAS) quantifies reproducibility across repeated executions of an aggregate feature ranking derived from local Monte-Carlo Shapley attributions. Global Relational Stability (GRS) measures cross-seed variation of a specified global score profile. The Structural Consistency Coefficient (SCC) quantifies ordinal agreement between the aggregate attribution ranking and a selected global ranking within a declared inspection scope. The metric names are retained for continuity, but "local" refers to the origin of the attributions; LAS does not measure per-instance explanation stability. Standard GRA is the primary instantiated global lens in this study for continuity with the originating G-XCI design and because its normalization, shared-extrema operator, and distinguishing coefficient can be specified and audited explicitly. It is treated as a conditional proxy rather than a privileged ground truth; at the protocol level, G-XCI accepts any declared global score vector whose construction and audit scope are specified [18–23].

The journal study expands the controlled evaluation in several directions while preserving a conservative interpretation of G-XCI. The matrix includes three tabular datasets and three model families—logistic regression, random forest, and XGBoost—with held-out predictive controls and explanation diagnostics. The primary GRA branch uses shared deviation extrema over the feature-by-sample comparison matrix, while the historical

per-feature-extrema implementation is retained only as an ablation. The global audit lens is varied through Standard GRA, the per-feature-extrema GRA variant, absolute Spearman-to-output association, and output permutation. SCC is stress-tested across inspection scales, GRA normalizations, distinguishing coefficients, and representation variants. Finally, the computational package uses code-fingerprinted checkpoints and completion hashes so that resumed computations cannot silently reuse results produced by different executed logic.

The resulting evidence changes the scientific emphasis of G-XCI. Adult Income does not exhibit a fixed SCC sign under the primary GRA contract; the direction changes with inspection scope. Alternative global audit lenses can produce substantially different SCC values for the same aggregate attribution ranking, and representation changes can materially alter SCC while held-out predictive performance changes little. A targeted WDBC audit further shows that complementing the binary output probability reverses the mean GRA-based SCC sign across all tested model–scale combinations while leaving the absolute local Shapley ranking unchanged. A second targeted audit forces Standard GRA, absolute Spearman-to-output, and output permutation onto the same 128 held-out cases and preserves all 27 original mean-SCC signs. These findings support an audit-contract interpretation rather than an intrinsic explanation-validity interpretation.

Accordingly, G-XCI's contribution is the audit architecture rather than a new predictive model or explanation algorithm. Unlike multi-metric benchmark suites such as OpenXAI, G-XCI keeps three questions separate—reproducibility of an aggregate local-attribution ranking (LAS), variation of a declared global score profile (GRS), and directional cross-view ordinal agreement within an explicit inspection scope (SCC)—and then subjects the audit contract itself to controlled sensitivity analysis. Output orientation, global lens, feature representation, audit sample scope, and inspection scale are therefore treated as disclosed components of the result rather than hidden implementation choices.

The present journal study makes five specific contributions:

- It formalizes G-XCI as a three-readout audit protocol that separates reproducibility of an aggregate ranking derived from local attributions (LAS), global-profile variation (GRS), and cross-view ordinal agreement (SCC), rather than collapsing explanation reliability into one truth score.
- It corrects and explicitly documents the Standard-GRA operator, isolates the legacy per-feature-extrema variant as an ablation, and evaluates alternative global audit lenses instead of assuming GRA is uniquely authoritative.
- It provides held-out, multi-dataset and multi-model evaluation together with sensitivity analyses over inspection scope, GRA normalization, the distinguishing coefficient, feature representation, output-class orientation, and matched audit-sample scope.
- It introduces an explicit audit-contract interpretation in which the attribution operator and baseline, target/output orientation, global proxy, preprocessing/representation, audit sample scope, repeated-execution design, and top-K scope are treated as part of the diagnostic result rather than hidden implementation details.
- It provides a reproducibility-oriented computational package with fresh computation, code fingerprints, disjoint representation calibration, tie-inclusive top-K handling, and file-level completion hashes.

The remainder of the article is organized as follows. Section 2 positions G-XCI relative to work on feature attribution, explanation stability, feature dependence, multi-metric evaluation, cross-view disagreement, and GRA. Section 3 defines the journal protocol and computational design. Section 4 reports the fresh-compute journal results. Section 5 discusses the implications and limitations of the conditional audit interpretation, followed by the conclusions.

## 2. Related Work

### 2.1. Local Feature Attribution and the Need for Explicit Evaluation

Feature-attribution methods provide a compact way to summarize how input variables contribute to individual predictions. SHAP derives additive attributions from Shapley values [1], LIME approximates the local decision surface with a weighted surrogate model [2], and Integrated Gradients attributes differentiable-model outputs along a path from a baseline to the observed input [3]. These methods answer different explanatory questions and rely on different assumptions; their outputs should therefore not be assumed interchangeable.

The XAI evaluation literature has repeatedly shown that explanations require their own validation criteria. Adebayo et al. demonstrated that visually convincing saliency outputs can be insensitive to model or data randomization, motivating explicit sanity checks [4]. Hooker et al. proposed quantitative benchmarking of feature-importance methods rather than relying on intuitive inspection [5]. Yeh et al. formalized infidelity and sensitivity as distinct explanation properties [6], while OpenXAI later systematized multiple quantitative metrics for faithfulness, stability, and fairness across methods and datasets [10]. More recently, Kelodjou et al. showed that Kernel SHAP can itself be unstable across repeated executions because of stochastic neighbor selection [11]. Taken together, these studies motivate a multi-dimensional view of explanation evaluation. G-XCI adopts the same general principle but asks a narrower tabular question about reproducibility of an aggregate ranking derived from local attributions, stability of a declared global profile, and rank agreement between the two views.

### 2.2. Stability, Reproducibility, and Uncertainty in Post Hoc Explanations

Instability has been documented as a practical weakness of perturbation-based post hoc explanations. Slack et al. showed that post hoc explanation methods can be manipulated adversarially, illustrating why attribution outputs should not be treated as unquestioned evidence [7]. Their later uncertainty-aware framework modeled variability in LIME and KernelSHAP explanations [8]. Zhou et al. developed S-LIME to stabilize LIME by controlling the perturbation sample size required for reproducible coefficients [9], and Kelodjou et al. directly analyzed stochastic instability in Kernel SHAP [11].

These studies are closely related to LAS, but the objectives are not identical. Prior stabilization methods primarily reduce or characterize stochastic variability of a particular explainer at the explanation level. In the present implementation, LAS is a diagnostic of the ranked vector obtained after aggregating absolute local Shapley attributions across a held-out audit subsample. It therefore measures end-to-end reproducibility of that aggregate ranking under the repeated-execution contract, which includes data resampling, model stochasticity when applicable, audit-subsample variation, and Monte-Carlo attribution randomness. It is not a guarantee of per-instance explanation stability or correctness.

Feature-ranking stability has also been formalized outside XAI. Wang et al. assessed repeated feature-importance rankings using pairwise Spearman correlations, and Nogueira et al. studied Spearman's rho specifically as a stability measure for feature rankings [24,25]. LAS follows this established same-view ranking-stability pattern, but applies it to aggregate absolute-Shapley rankings under the repeated-execution contract.

### 2.3. Feature Dependence, Preprocessing, and Representation

The semantics of feature attribution depend on the feature distribution and representation. Aas et al. showed that Shapley explanations can change materially when feature dependence is modeled rather than ignored, especially for correlated predictors [12]. Kumar et al. further emphasized that Shapley-based feature-importance claims depend

on how the cooperative game and missing-feature semantics are defined [13]. These results caution against interpreting an attribution vector independently of its background distribution and feature representation.

Recent empirical work on tabular explanation accuracy provides a complementary perspective. Rahnama et al. found that explanation performance can depend on predictor correlation, the mixture of categorical and numerical variables, preprocessing, sample size, and the model being explained [14]. Applied studies have likewise shown that data issues, encoding choices, and collinearity can influence the usefulness or consistency of post hoc explanations [15,16]. These findings motivate the preprocessing-aware design of G-XCI and the explicit E6 representation audit. Rather than claiming that one-hot encoding or correlation filtering solves feature dependence, G-XCI treats the representation contract as part of what must be disclosed and stress-tested.

### 2.4. Cross-View Agreement and Multi-Metric Explanation Auditing

Most explanation metrics evaluate one property of one explainer. However, disagreement between distinct explanatory views can itself be informative. Mothilal et al. compared feature-attribution explanations with counterfactual explanations and reported that top-ranked features from SHAP or LIME need not agree with features that are necessary or sufficient under a counterfactual criterion [17]. This result does not establish one view as superior; it demonstrates that different explanation paradigms operationalize different questions.

G-XCI extends this cross-view idea in a deliberately modest way. It does not attempt to unify the semantics of the two views. SCC asks whether an aggregate ranking derived from local attributions and a separately defined global ranking are ordinally compatible within a declared inspection scope. Kendall's tau-b is used because the comparison is rank-based and ties can occur [26]. This separates cross-view agreement from aggregate-attribution reproducibility and global-profile variation. G-XCI therefore complements multi-metric frameworks such as OpenXAI [10] rather than replacing them. OpenXAI separates evaluation properties such as faithfulness, stability/robustness, and fairness; G-XCI asks a narrower question by separating repeated-run ost hocor within each declared view from conditional ordinal compatibility between explicitly specified views.

### 2.5. Grey Relational Analysis as a Conditional Global Relational Lens

Grey-system theory was developed to analyze systems characterized by incomplete or partially known information [18]. Modern grey-system methodology includes multiple grey-incidence and relational-analysis formulations rather than one universal notion of importance [19]. In a conventional GRA workflow, normalized comparison sequences are converted to absolute deviations from a reference sequence, and the grey relational coefficient is then calculated using minimum and maximum deviations shared over the comparison domain. Peer-reviewed formulations explicitly express these extrema as nested minima and maxima over sequences and observation indices [21,22]. This is the operator adopted as Standard GRA in the journal branch of G-XCI.

Two methodological consequences are particularly relevant. First, normalization is part of the GRA construction because relational coefficients are computed after sequence transformation; the resulting ranking is therefore conditional on the declared normalization contract [22]. Second, the distinguishing coefficient is not guaranteed to be innocuous. Mahmoudi et al. demonstrated that changing the distinguishing coefficient can change GRA rankings and recommended sensitivity analysis rather than assuming the conventional value of 0.5 is universally harmless [20]. These considerations motivate the E5 ablations over normalization and zeta in the present study.

This matters because a negative SCC indicates ordinal disagreement between the specified local and global rankings; it does not reveal which ranking is correct. The journal study tests this interpretation by replacing Standard GRA with absolute Spearman-to-output association and output permutation, allowing the sensitivity of SCC to the declared global lens to be observed directly.

### 2.6. Positioning and Research Gap Addressed by G-XCI

The literature above establishes several necessary ingredients for reliable explanation practice: feature-attribution methods require explicit evaluation [4–6], post hoc outputs can be unstable [7–9,11], feature dependence and preprocessing affect attribution semantics [12–16], different explanation paradigms can disagree [17], and GRA is conditional on operator and preprocessing choices [18–23]. These strands do not imply a missing universal metric. They leave a narrower engineering gap: practitioners may need to assess whether an aggregate attribution ranking is repeatable, whether a separately specified global view has low repeated-run variation, and whether the two orderings agree while making the representation, target orientation, audit sample scope, proxy, and inspection scope explicit.

G-XCI is positioned as a response to that gap. LAS uses mean pairwise Spearman rank correlation to summarize reproducibility of the aggregate ranking derived from local attributions, following established feature-ranking stability practice [24,25,27]. GRS summarizes cross-seed variance of the chosen global score profile. SCC uses Kendall tau-b to summarize cross-view ordinal agreement under a tie-inclusive top-K policy [26]. The journal experiment then subjects the protocol itself to sensitivity analysis over model family, global proxy, inspection scope, GRA operator, normalization, distinguishing coefficient, and feature representation, while predictive behavior is reported separately. This design makes major audit assumptions visible rather than treating them as hidden implementation details.

The resulting positioning is therefore complementary to both explanation methods and explanation benchmarks. G-XCI does not replace SHAP, LIME, Integrated Gradients, OpenXAI, or dependence-aware attribution methods. It also does not claim that GRA is the correct population-level explanation. Its role is to expose when reliability conclusions are conditional on choices in the audit pipeline and to provide a reproducible protocol for investigating those dependencies. Table 1 summarizes this positioning relative to the principal related-work strands.

**Table 1.** Positioning of G-XCI relative to the principal related-work strands.

| Related-Work Strand | Primary Evaluation Target | Representative Anchors | Relation to G-XCI |
| --- | --- | --- | --- |
| Attribution foundations | Local feature contribution semantics | [1–3] | Provides candidate local views; this study uses Monte-Carlo Shapley as the local operator. |
| Explanation evaluation and benchmarking | Faithfulness, sensitivity, sanity, multi-metric evaluation | [4–6,10] | Motivates explicit evaluation; G-XCI narrows the focus to reproducibility, global stability, and cross-view order agreement. |
| Stability and uncertainty | Run-to-run variability of local explanations | [7–9,11] | Contextualizes stochastic instability; LAS reports reproducibility of an aggregate ranking derived from local attributions. |
| Dependence and preprocessing | Effect of correlation, encoding, and feature representation | [12–16] | Motivates explicit preprocessing/representation contracts and E6 stress tests. |
| Cross-view explanation comparison | Agreement versus disagreement between different explanatory criteria | [17] | SCC quantifies ordinal compatibility without declaring either view ground truth. |
| Grey relational analysis | Relational similarity under specified normalization/operator choices | [18–22] | Provides one conditional global lens; Standard/Legacy definitions and zeta/normalization are audited explicitly. |

In summary, the literature supports explicit explanation evaluation but not a universal interpretation of any single attribution or global-importance vector. G-XCI therefore adopts a conditional audit stance: every LAS, GRS, and SCC result is interpreted with respect to the stated model, local attribution operator and baseline, target/output orientation, global audit lens, preprocessing/representation contract, audit sample scope, repeated-execution design, and inspection scope. This framing provides the conceptual bridge to Section 3.

## 3. Materials and Methods

All journal computations were executed in Python 3.13.15 with NumPy 2.1.3, pandas 2.2.3, SciPy 1.16.3, scikit-learn 1.6.1, and XGBoost 3.4.1; the exact software environment is archived in Supplementary File S1.

This study evaluates G-XCI as a post hoc, multi-view audit protocol for tabular explanation reliability. The journal protocol was designed to separate three questions that are often conflated in explanation analysis: whether an aggregate ranking derived from local attributions is reproducible across repeated executions, whether a specified global relational profile is stable across the same executions, and whether the two ranked views agree within a declared inspection scope. G-XCI therefore reports three complementary readouts—Local Attribution Stability (LAS), Global Relational Stability (GRS), and the Structural Consistency Coefficient (SCC)—without treating any single local or global view as causal ground truth. This distinction follows the broader XAI literature calling for explicit and multi-dimensional evaluation of explanation behavior rather than reliance on visually plausible single-run explanations [4–10].

### *3.1. Study Design and Repeated-Execution Protocol*

The final journal experiment used five repeated seeds, S = 5, with values {42, 101, 202, 303, 404}. For each dataset–model pair, a fresh stratified 70/30 train/test split was constructed separately for every seed. The same seed also controlled model stochasticity when applicable, Monte-Carlo Shapley sampling, and the label-free permutation proxy. Consequently, LAS and the seed-level SCC summaries quantify end-to-end reproducibility under the repeated-execution design, which includes resampling and model/explainer randomness rather than isolating explainer stochasticity alone.

The principal journal matrix used the CALIBRATED Monte-Carlo profile: 32 feature-order permutations and 128 held-out evaluation samples per seed. These values are fixed computational budgets rather than claimed optima. The primary inspection scope was K = 20, with K = 10 and the full feature set used in the scale-of-inspection analysis. Standard GRA used $\zeta = 0.50$ by default, with $\zeta \in \{0.25, 0.50, 0.75\}$ evaluated in the GRA sensitivity study.

### *3.2. Datasets and Target Coding*

Three binary tabular datasets were used to create contrasting sample-size, dimensionality, and representation regimes. IBM HR Attrition and UCI Adult Income were loaded from the frozen source package inherited from the historical pipeline, whereas Wisconsin Diagnostic Breast Cancer (WDBC) was loaded through scikit-learn. Targets were converted to {0,1} by explicit mapping. For WDBC, the scikit-learn coding was retained (target = 1 denotes benign). Because GRA compares feature sequences with the probability of the selected class, output-class orientation is part of the audit contract; the WDBC analysis therefore has no clinical interpretation, and its SCC sign is not presented as class-orientation invariant.

The Adult Income dataset is publicly available from the UCI Machine Learning Repository [28]. The IBM HR Analytics Employee Attrition & Performance dataset is publicly available through Kaggle [29]. The Wisconsin Diagnostic Breast Cancer dataset is avail-

able from the UCI Machine Learning Repository and through scikit-learn [30]. The reproducibility package and the targeted supplementary audit outputs (Tables S1 and S2, Figures S1 and S2, executed notebook, run manifest, and completion receipt) are provided as Supplementary Materials with the submission. No new private participant data were collected. Dataset characteristics under the fixed schema are summarized in Table 2.

**Table 2.** Dataset characteristics under the fixed feature-schema contract.

| Dataset | N | Raw Predictors | Positive Class | Positive N | Final Encoded D |
|---|---|---|---|---|---|
| IBM HR Attrition | 1470 | 34 | Attrition = Yes | 237 | 52 |
| UCI Adult Income | 48,842 | 14 | Income > 50 K | 11,687 | 108 |
| WDBC | 569 | 30 | scikit-learn target = 1 (benign) | 357 | 30 |

### 3.3. Fixed Feature Schema and Preprocessing

A fixed-schema preprocessor was used so that feature ranks could be compared across seeds without dimensionality drift. The representation schema was defined from predictor values only and never used the target. Raw numeric predictors with at most one observed non-missing value and raw categorical predictors with at most one observed category were removed as dataset-level constants. For IBM HR, this step removed EmployeeCount, Over18, and StandardHours; the encoded dimensionality changed from 55 before constant removal to 52 final predictors. Adult Income and WDBC had no dataset-level constant raw predictors.

For nonconstant categorical variables, the complete dataset-level predictor support was used to fix a sorted category vocabulary and full one-hot feature names. Missing categorical values were represented by a dedicated "__MISSING__" level. Numerical imputation parameters were not learned globally: each repeated training split supplied its own numeric median, and the fitted median was then applied to that seed's held-out test split. Thus, outcome labels and learned numeric/model parameters remained split-specific, whereas the rank space itself was deliberately fixed at the target-free dataset level.

This target-free schema choice is methodological rather than a claim of strict train-only representation discovery. Accordingly, the E2–E5 predictive controls should be read under this same target-free transductive schema contract rather than as strict deployment-performance estimates. It was adopted because LAS and SCC require aligned feature identities across repeated runs. The Discussion treats this dataset-level predictor access as an explicit protocol condition and identifies train-only vocabulary estimation followed by post hoc feature-space alignment as a future sensitivity analysis.

### 3.4. Predictive Models

Three fixed model families were evaluated without hyperparameter optimization: logistic regression, random forest, and XGBoost. This design broadens model-family coverage while avoiding a nested model-selection process that could obscure the explanation-audit comparison. Logistic regression was trained inside a StandardScaler pipeline. Tree models used unscaled encoded features. The fixed model configurations are listed in Table 3.

**Table 3.** Fixed predictive-model configurations used in the journal experiment.

| Model | Fixed Configuration |
|---|---|
| Logistic Regression | StandardScaler → LogisticRegression; solver = lbfgs; max_iter = 3000; random_state = seed. |
| Random Forest | 300 trees; min_samples_leaf = 2; max_features = sqrt; random_state = seed; n_jobs = 2. |
| XGBoost | 300 trees; max_depth = 4; learning_rate = 0.05; subsample = 0.90; colsample_bytree = 0.90; reg_lambda = 1.0; objective = binary:logistic; eval_metric = logloss; random_state = seed; n_jobs = 2. |

Class predictions used a fixed probability threshold of 0.5. Accuracy, balanced accuracy, and F1 were computed from these decisions, while ROC-AUC and average precision were computed from held-out predicted probabilities. These predictive quantities are controls for model behavior and are not used to declare an explanation view correct.

### 3.5. Monte-Carlo Shapley Local Attribution View

The attribution view is a model-agnostic Monte-Carlo permutation-Shapley approximation rather than an exact invocation of a specific SHAP software explainer. It is a single-baseline coalition construction and does not model the conditional feature distribution or enforce grouped coalitions for one-hot blocks. Shapley-based attribution provides a common additive feature-attribution vocabulary [1], while prior work cautions that its semantics depend on the feature game, dependence assumptions, representation, and background construction [12,13]. In this study, local attributions are computed for individual held-out cases and then aggregated into one feature-ranking vector per repeated execution; consequently, the subsequent LAS and SCC analyses do not evaluate per-instance explanation stability.

For each seed, the training-set feature-wise median formed the baseline vector $\mu$. From the held-out evaluation split, $M = 128$ cases were selected without replacement. For every selected case and each of $P = 32$ random feature orders, the algorithm began at $\mu$ and sequentially replaced baseline coordinates with the case's observed values. Marginal changes in the model's selected-class probability were assigned to the introduced feature. Contributions were averaged over permutations within each case, converted to absolute values, and then averaged over the $M$ audited cases to form the aggregate attribution vector:

$$\bar{\phi}(s,j) = \frac{1}{M}\sum_{m=1}^{M}\left|\frac{1}{P}\sum_{p=1}^{P}\Delta f(s,m,p,j)\right|. \quad (1)$$

The vector $\bar{\phi}(s)$ was ranked in descending order using average ranks for exact ties. The term "local" in LAS refers to the origin of these case-level attributions; operationally, the ranked object used by LAS and SCC is the aggregate vector $\bar{\phi}$(s). The coordinate-wise training-median baseline was fixed across datasets, models, and representations so that E2–E6 share the same local-attribution construction; it is not claimed to be dependence-aware. For one-hot categorical blocks, a coordinate-wise median can be off-manifold (for example, an all-zero dummy group), so the resulting aggregate ranking is explicitly baseline-conditioned. E6 varies the encoded representation under this fixed attribution construction; it does not substitute for dependence-aware or grouped-feature attribution. Batching in groups of at most four evaluated cases was used only to reduce peak memory and does not change the attribution semantics.

### 3.6. Local Attribution Stability (LAS)

Let rL(s) denote the descending rank vector of the aggregate attribution scores $\bar{\phi}$(s) for repeated execution s. LAS is the mean pairwise Spearman rank correlation over all distinct seed pairs [27]:

$$\mathrm{LAS} = [2/(\mathrm{S}(\mathrm{S}-1))]\Sigma[\mathrm{s} < \mathrm{t}]\rho\mathrm{S}(\mathrm{rL}(\mathrm{s}), \mathrm{rL}(\mathrm{t})) \quad (2)$$

LAS follows the same-view feature-ranking stability pattern in which agreement among repeated rankings is summarized by pairwise Spearman correlation [24,25]. It is a reproducibility statistic for the seed-specific aggregate attribution ranking, not a measure of individual case-level explanation stability. Because the aggregate vector in Equation (1) averages absolute contributions over $M = 128$ held-out cases before ranking, a high LAS can coexist with substantial instance-level attribution variability. High LAS also does not establish faithfulness, causality, or agreement with another explanatory view.

### 3.7. Grey Relational Global View

#### 3.7.1. Sequence Normalization

For each held-out execution, every encoded feature sequence and the model's held-out selected-class probability sequence were normalized independently before Grey Relational Analysis (GRA). The primary compatibility path retained a robust-z transformation followed immediately by min–max normalization; E5 additionally evaluated direct min–max, 1% winsorization followed by min–max, and rank-ECDF normalization. Because the reference sequence is the selected-class probability, GRA is conditional on output orientation as well as on normalization and feature representation [19–23].

$$z_i = (a_i - \text{median}(a))/[1.4826 \times \text{MAD}(a)] \quad (3)$$

When MAD was smaller than $\epsilon = 10^{-12}$, the implementation used $z_i = a_i - \text{median}(a)$. Constant sequences were mapped to zeros by the subsequent min–max routine.

**Lemma 1.** *Affine equivalence of robust-z followed by min–max.*

For any nonconstant sequence a and any positive affine transformation z_i = (a_i − m)/s with s > 0, min–max normalization of z is exactly the same as min–max normalization of a. This follows because min(z) = [min(a) − m]/s and max(z) = [max(a) − m]/s, so the shift m and scale s cancel:

$$[z_i - \min(z)] \,/\, [\max(z) - \min(z)] = [a_i - \min(a)] \,/\, [\max(a) - \min(a)] \quad (4)$$

The E5 experiment therefore treats the observed $\sim 10^{-16}$ difference between the two implementations as a floating-point verification of the code path, not as empirical evidence that robust-z adds a distinct robustification effect before min–max.

#### 3.7.2. Standard GRA and the Legacy Ablation

Let $\tilde{x}_ij$ be the normalized value of feature j for held-out case i and $\tilde{p}_i$ the independently normalized model-output sequence. The feature-by-sample deviation matrix is

$$\Delta_ij = | \tilde{x}_ij - \tilde{p}_i |. \quad (5)$$

The primary Standard-GRA operator uses one shared minimum and maximum over the complete deviation matrix within an execution:

$$\Delta_min = \min(i,j)\ \Delta_ij,\ \Delta_max = \max(i,j)\ \Delta_ij. \quad (6)$$

This shared-extrema construction matches published GRA formulations that define the extrema across comparison sequences and observation indices [21,22]. The grey relational coefficient and feature-level relational grade are

$$\xi_ij = (\Delta_min + \zeta\Delta_max) \,/\, (\Delta_ij + \zeta\Delta_max + \epsilon),\ \gamma_j = (1/n)\ \Sigma[i = 1..n]\ \xi_ij, \quad (7)$$

with $\epsilon = 10^{-12}$ for numerical protection and $\zeta = 0.50$ by default. The conference-era per-feature implementation was not deleted; it was isolated as a legacy ablation in which $\Delta_min,j$ and $\Delta_max,j$ are computed separately for each feature. E5 directly compares the Standard and Legacy operators so that the journal version does not retroactively rewrite the historical computation.

The distinguishing coefficient $\zeta$ was varied over 0.25, 0.50, and 0.75 because published evidence shows that $\zeta$ can affect GRA rankings and should not automatically be treated as irrelevant [20].

### 3.8. Alternative Global Audit Lenses

The global-lens analysis compared four score vectors: Standard GRA, the historical per-feature-extrema GRA variant, absolute Spearman-to-output, and a label-free output-permutation proxy. Standard GRA, the per-feature GRA variant, and absolute Spearman were computed on the complete held-out split. The permutation proxy used the complete held-out split when it contained at most 4096 rows and otherwise a seeded subsample of 4096 rows. The absolute Spearman proxy assigns feature j the score $|\rho S(X_j,p)|$, with constant features assigned zero. The output-permutation proxy repeatedly permutes one held-out feature at a time and records the mean absolute change in model probability:

$$PI_j = (1/R)\ \Sigma[r = 1..R]\ mean_i\ |\ p(x_i) - p(x_i \text{ with feature j permuted in repeat r})\ |, \quad (8)$$

where R = 5. The permutation proxy is model-output-based and label-free; it is conceptually related to permutation feature importance [31] but the exact score used here is the mean absolute probability change rather than a loss increase. These proxies are alternative audit lenses, not external ground-truth feature importance. The audit sample scope is therefore also explicit: the aggregate Shapley ranking uses M = 128 held-out cases, GRA and absolute Spearman use the complete held-out split, and output permutation uses up to 4096 cases. Cross-view SCC is interpreted conditional on these stated scopes rather than as a pure comparison of operators on an identical case set.

### 3.9. Global Relational Stability (GRS)

For a chosen global score vector g(s) = (g(s,1), . . . , g(s,d)) across S repeated executions, GRS is the mean feature-wise sample variance:

$$GRS = (1/d)\ \Sigma[j = 1..d]\ Var_s(\ g(s,j)\ ). \quad (9)$$

Lower GRS indicates that the numerical global profile varies less across repeated executions. GRS is computed on score values, whereas E3 and E5 additionally report the pairwise Spearman stability of the corresponding global rank vectors as an auxiliary diagnostic. Because GRS is a variance on the native score scale, its magnitude is interpreted only within a fixed global-score definition and normalization contract; raw GRS values are not compared as if they were commensurate across different audit lenses.

### 3.10. Structural Consistency Coefficient (SCC) and Tie-Inclusive Top-K

SCC measures ordinal agreement between the aggregate attribution ranking rL(s) and a specified global ranking. Kendall's $\tau$_b was used because tied ranks can occur [26]. For each seed, the comparison subset is selected from rL(s), not from the global ranking. If the nominal K-th aggregate-attribution rank is tied, every feature sharing that boundary rank is included, so the effective K can exceed nominal K. This policy avoids arbitrary dependence on original column order.

$$SCC(s,K) = \tau_b(\ rL(s)\,|\,T(s,K),\ rG(s)\,|\,T(s,K)\ ),\ SCC(K) = (1/S)\ \Sigma[s = 1..S]\ SCC(s,K), \quad (10)$$

where T(s,K) is the tie-inclusive subset selected from the aggregate attribution ranking. K = 20 is the primary inspection scope, with K = 10 and all features used in the scale analysis. The phrase "scale of inspection" is an operational interpretation introduced in this study: K specifies how much of the ranked feature space is compared. SCC is therefore conditional

on the global lens, target orientation, audit sample scope, and K; it is not a correctness or causal-validity score. For nominal top-K analyses, the subset is selected from the aggregate local-attribution ranking, so this is a directional audit of how the global ranking orders the locally selected features; the all-feature SCC removes this subset-selection asymmetry.

### 3.11. Experimental Modules

E1 hash-verifies the frozen MLMI/ICKII provenance artifacts; E0 audits constant removal and fixed feature dimensions; E2 is the primary 3-dataset × 3-model evaluation at Standard GRA and K = 20 with held-out predictive controls; and E3 compares Standard GRA, Legacy GRA, absolute Spearman-to-output, and output permutation.

E4 evaluates K = 10, K = 20, and all-feature inspection scopes; E5 evaluates Standard versus Legacy extrema, $\zeta \in \{0.25, 0.50, 0.75\}$, four normalization modes, and the affine-equivalence implementation check; and E6 evaluates Original, DropFirstCategorical, and HighCorrFilter representations under logistic regression with disjoint target-free calibration.

### 3.12. E6 Disjoint Representation Calibration

E6 was designed specifically to prevent the correlation-based representation filter from being fit on observations that later appear in E6 training or testing. A deterministic, target-free 20% representation-calibration subset was drawn once per dataset using the deterministic seed identifier `20260827`; the remaining 80% formed the E6 evaluation pool. No target values from the calibration subset were used. The calibration/evaluation sizes were 294/1176 for IBM HR, 9768/39,074 for Adult, and 114/455 for WDBC.

Within the calibration subset, the same fixed target-free feature schema was applied, while numeric medians for the calibration transformation were learned from the calibration data itself. Three representations were defined. Original retained the full encoded feature set. DropFirstCategorical removed the first sorted category indicator from every categorical predictor. HighCorrFilter first removed calibration-constant encoded features and then applied a deterministic greedy filter to pairs with absolute Spearman correlation $|\rho| \geq 0.95$: pairs were sorted by decreasing correlation, and the second feature was dropped when both members remained available. The resulting feature subset was then frozen and reused across all five E6 seeds so that LAS and SCC remained defined in a common rank space.

For each E6 representation, only the disjoint 80% evaluation pool was subsequently split 70/30 by seed, and only logistic regression was fitted. Thus, the E6 correlation filter had neither label access nor covariate overlap with the E6 repeated train/test samples.

### 3.13. Descriptive Seed Uncertainty and Statistical Interpretation

For SCC, the mean and sample standard deviation across the five seed-specific values were reported. A nonparametric bootstrap with 5000 resamples of the five observed seed values generated a 95% descriptive resampling interval. Because the five executions are not a population sample and their train/test splits overlap in observations, these intervals are used only as a compact summary of variation within the specified repeated-execution design. They are not population confidence intervals, null-hypothesis tests, or evidence of statistical significance.

### 3.14. Use of Generative AI-Assisted Research Tools

During study and manuscript preparation, the authors used ChatGPT (OpenAI; https://chatgpt.com/; accessed on 18 September 2026), Claude (Anthropic; https://claude.ai/; accessed on 18 September 2026), Google Gemini (https://gemini.google.com/; accessed on 18 September 2026), Microsoft Copilot (https://copilot.microsoft.com/; accessed on 18 September 2026), SciSpace (https://scispace.com/; accessed on 18 September 2026), and Perplexity Pro (Perplexity AI; https://www.perplexity.ai/; accessed on 18 September 2026).

Their roles were limited to adversarial code review, methodological cross-checking, candidate literature discovery and verification, and manuscript-language/formatting assistance. These tools did not autonomously generate or replace the reported experimental data or numerical results. Candidate bibliographic information was checked against publisher, proceedings, repository, or DOI records before inclusion; code changes and numerical claims were accepted only after execution or artifact-level verification by the authors. The authors reviewed and edited all AI-assisted output and remain fully responsible for the study design, analysis, interpretation, and manuscript content.

### *3.15. Computational Provenance and Reproducibility*

The final journal implementation used streaming, seed-level checkpoints to reduce memory pressure and permit recovery after runtime disconnection. A checkpoint was accepted only when its checkpoint schema, run-configuration hash, and runtime computation-code SHA-256 matched the current execution. The E6 cache additionally incorporated a SHA-256 fingerprint of the realized calibration/evaluation indices and representation feature subsets. The final run recomputed all 45 E2–E5 jobs and all 45 E6 jobs from the code-fingerprinted namespace rather than resuming older payloads.

The completion receipt stores the run-configuration hash, E2–E5 and E6 computation-code fingerprints, the E6 filter-specification fingerprint, cache statistics, and SHA-256 hashes for all required exported tables and figures. Historical MLMI/ICKII artifacts are maintained as frozen provenance and hash-verified against their reference files; they are not claimed to have been recomputed in the final journal run. This separation preserves the historical record while making the fresh-compute journal outputs the sole numerical authority for the journal manuscript.

### *3.16. Targeted Supplementary Audits of Output Orientation and Audit Sample Scope*

Two targeted supplementary audits were conducted after the primary journal computation was frozen. They used the same five seeds, 70/30 stratified splitting rule, three model families, 32-permutation $\times$ 128-case Monte-Carlo Shapley budget, Standard GRA with shared extrema, $\zeta = 0.5$, and the same software environment as the frozen journal run. The supplementary run independently computed 45 dataset–model–seed jobs, with zero checkpoint resumes, and verified matching source-package, environment, and scientific-configuration fingerprints against the frozen main experiment.

Supplementary Audit S1 evaluates WDBC output orientation. In scikit-learn, target 1 denotes benign, and target 0 denotes malignant. For each seed and model, the benign probability p and its complement $1-p$ were compared as GRA reference sequences at K = 10, K = 20, and all features. To isolate the GRA output-orientation effect, the primary S1 comparison fixes the class-1 aggregate absolute-Shapley ranking for both orientations. A secondary control independently recomputes class-0 Shapley attributions using the identical held-out audit cases and feature-order permutation stream, allowing direct verification that the absolute local ranking is unchanged under binary probability complementation.

Supplementary Audit S2 evaluates audit-sample-scope confounding in the global-lens comparison. For every dataset, model, and seed, Standard GRA, absolute Spearman-to-output, and output permutation are recomputed on exactly the same 128 held-out cases selected by the Monte-Carlo Shapley audit routine. The resulting SCC values are compared with the original E3 values, in which the global lenses used their protocol-specific held-out scopes. S2 therefore tests whether the qualitative lens-disagreement pattern survives when audit sample scope is held fixed.

## 4. Results

All journal-extension results reported in this section are taken exclusively from the code-fingerprinted fresh journal computation. The historical MLMI/ICKII tables are retained only as frozen provenance and are not used as the numerical basis for the journal claims below. Predictive metrics are reported as held-out controls; they are not treated as evidence that a particular explanation view is correct.

### 4.1. Multi-Dataset and Multi-Model Diagnostic Profiles

Table 4 summarizes the primary Standard-GRA results at nominal K = 20 across three datasets and three model families. The 45 individual seed-level E2 results underlying these five-seed summaries are reported in Supplementary Table S3, copied directly from the frozen authoritative `tableE2_seed_level.csv`; no additional computation was introduced for that display. LAS ranged from 0.756 to 0.949, showing that reproducibility of the aggregate attribution ranking was high in some settings but not uniform. GRS ranged from approximately $2 \times 10^{-6}$ to $2.1 \times 10^{-4}$ on the GRA score scale. Because no externally validated GRS threshold is assumed, these values are reported descriptively rather than labeled "stable" or "unstable". SCC showed materially different signs and magnitudes across datasets and models.

For IBM HR Attrition, mean SCC was −0.107 for logistic regression, +0.074 for random forest, and −0.078 for XGBoost. The corresponding 95% descriptive five-seed resampling intervals were [−0.162, −0.044], [0.015, 0.133], and [−0.179, 0.021]. Adult Income showed positive mean SCC at K = 20 for all three model families: +0.063, +0.135, and +0.122, with descriptive intervals [−0.084, 0.164], [0.072, 0.198], and [−0.004, 0.236]. Under the retained WDBC class-1 (benign) orientation, mean SCC was negative for logistic regression (−0.261), random forest (−0.495), and XGBoost (−0.322). These intervals are descriptive summaries only and are not used for significance claims.

Held-out predictive performance differed across model–dataset pairs but did not map monotonically onto SCC. Adult Income XGBoost achieved the highest Adult ROC-AUC (0.926) and average precision (0.825), while its Standard-GRA SCC remained modestly positive (+0.122). On WDBC, all three models achieved ROC-AUC values above 0.987, yet all three exhibited negative Standard-GRA SCC. These observations motivate treating predictive performance and explanation agreement as separate diagnostic dimensions.

**Table 4.** Primary journal diagnostic profile at K = 20 (Standard GRA).

| Dataset | Model | LAS | GRS | SCC | Desc. 5-Seed Interval | ROC-AUC | Avg. Precision |
|---|---|---|---|---|---|---|---|
| IBM HR Attrition | Logistic Regression | 0.756 | $1.15 \times 10^{-4}$ | −0.107 | [−0.162, −0.044] | 0.830 | 0.596 |
| IBM HR Attrition | Random Forest | 0.949 | $1.24 \times 10^{-4}$ | 0.074 | [0.015, 0.133] | 0.814 | 0.559 |
| IBM HR Attrition | XGBoost | 0.905 | $7.61 \times 10^{-5}$ | −0.078 | [−0.179, 0.021] | 0.810 | 0.575 |
| UCI Adult Income | Logistic Regression | 0.883 | $3.26 \times 10^{-6}$ | 0.063 | [−0.084, 0.164] | 0.905 | 0.765 |
| UCI Adult Income | Random Forest | 0.909 | $2.58 \times 10^{-6}$ | 0.135 | [0.072, 0.198] | 0.917 | 0.801 |
| UCI Adult Income | XGBoost | 0.891 | $2.26 \times 10^{-6}$ | 0.122 | [−0.004, 0.236] | 0.926 | 0.825 |
| WDBC | Logistic Regression | 0.783 | $1.43 \times 10^{-4}$ | −0.261 | [−0.398, −0.124] | 0.992 | 0.995 |
| WDBC | Random Forest | 0.931 | $1.62 \times 10^{-4}$ | −0.495 | [−0.560, −0.429] | 0.988 | 0.988 |
| WDBC | XGBoost | 0.819 | $2.10 \times 10^{-4}$ | −0.322 | [−0.360, −0.284] | 0.993 | 0.995 |

### 4.2. Dependence on the Global Audit Lens

The alternative-proxy analysis showed that cross-view rank agreement was strongly conditioned on the chosen global audit lens. Across the nine dataset–model combinations, the output-permutation proxy produced positive SCC values from 0.669 to 0.886, and absolute Spearman-to-output was also positive in all cases (0.192 to 0.680). Standard GRA

produced mixed signs for IBM HR and Adult and negative values for WDBC under the retained benign-class orientation. A targeted shared-scope audit subsequently recomputed Standard GRA, absolute Spearman, and output permutation on exactly the same 128 held-out cases used by the Shapley routine. All 27 dataset–model–lens mean-SCC signs were preserved; the largest absolute change in SCC was 0.095. Thus, audit sample scope affects magnitude, but the qualitative global-lens disagreement observed in E3 is not explained by unequal held-out sample scopes.

Several WDBC settings illustrate the magnitude of lens sensitivity under the retained output orientation. For random forest, SCC was −0.495 under Standard GRA and −0.354 under the per-feature-extrema GRA variant, but +0.680 under absolute Spearman-to-output and +0.886 under output permutation. For logistic regression on IBM HR, the same aggregate attribution ranking yielded SCC = −0.107 under Standard GRA, +0.225 under absolute Spearman-to-output, and +0.819 under output permutation. SCC here reflects agreement with the specified global lens (Section 3.10). Figure 1 provides a compact visual summary of this audit-lens dependence across all nine dataset–model combinations. The corresponding numerical values are reported in Table 5.

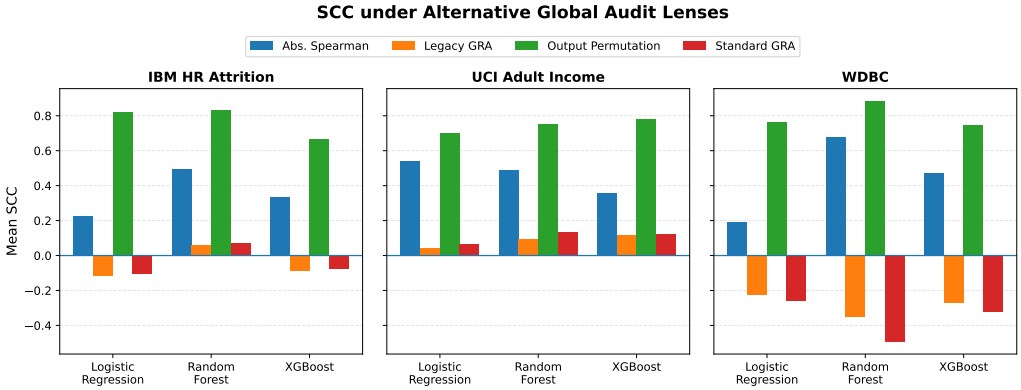



**Figure 1.** Cross-view SCC under alternative global audit lenses at nominal $K = 20$. The y-axis is the five-seed mean SCC, and the panels separate the three datasets; within each panel, bars compare model families and declared global lenses. Standard GRA uses shared global deviation extrema, whereas the legacy branch uses per-feature extrema. The contrast is descriptive and illustrates audit-lens dependence rather than explanation correctness.

**Table 5.** SCC under alternative global audit lenses (K = 20).

| Dataset | Model | Standard GRA | Legacy GRA | Abs-Spearman | Output Permutation |
|---|---|---|---|---|---|
| IBM HR Attrition | Logistic Regression | −0.107 | −0.116 | 0.225 | 0.819 |
| IBM HR Attrition | Random Forest | 0.074 | 0.059 | 0.495 | 0.832 |
| IBM HR Attrition | XGBoost | −0.078 | −0.088 | 0.336 | 0.669 |
| UCI Adult Income | Logistic Regression | 0.063 | 0.042 | 0.541 | 0.701 |
| UCI Adult Income | Random Forest | 0.135 | 0.093 | 0.487 | 0.752 |
| UCI Adult Income | XGBoost | 0.122 | 0.118 | 0.358 | 0.779 |
| WDBC | Logistic Regression | −0.261 | −0.223 | 0.192 | 0.764 |
| WDBC | Random Forest | −0.495 | −0.354 | 0.680 | 0.886 |
| WDBC | XGBoost | −0.322 | −0.274 | 0.469 | 0.749 |

### *4.3. Scale-of-Inspection Sensitivity*

Changing the inspection scope K materially altered SCC in several settings. Adult Income provided the clearest example of sign reversal across scale. For logistic regression, SCC changed from −0.191 at K = 10 to +0.063 at K = 20 and then to −0.474 when all 108 encoded predictors were included. The corresponding sequence for random forest was −0.173, +0.135, and −0.495, and for XGBoost it was −0.120, +0.122, and −0.491. Thus,

the corrected journal protocol does not support the earlier conference-era description of Adult as persistently negative across top-K settings; instead, Adult exhibits pronounced scale-of-inspection dependence.

Under the retained WDBC class-1 (benign) orientation, SCC was negative for all three models at K = 10, K = 20, and the full feature set. This is an orientation-conditional result rather than an invariant property of the dataset because the GRA reference sequence is the selected-class probability. The tie-inclusive selection policy did not alter any reported value because BoundaryTie_Fraction was 0.0 throughout E2–E6; effective K equaled nominal K in every tested configuration. Figure 2 summarizes the corresponding scale-of-inspection pattern across the full model–dataset matrix. The corresponding numerical values are reported in Table 6.

Because $T(s, K)$ is selected from the local attribution ranking, nominal top-$K$ SCC answers a directional question: among the features prioritized by the local view, how similarly does the declared global view order those same features? It is not a symmetric comparison of independently selected top-$K$ sets. The all-feature SCC removes this subset-selection asymmetry.

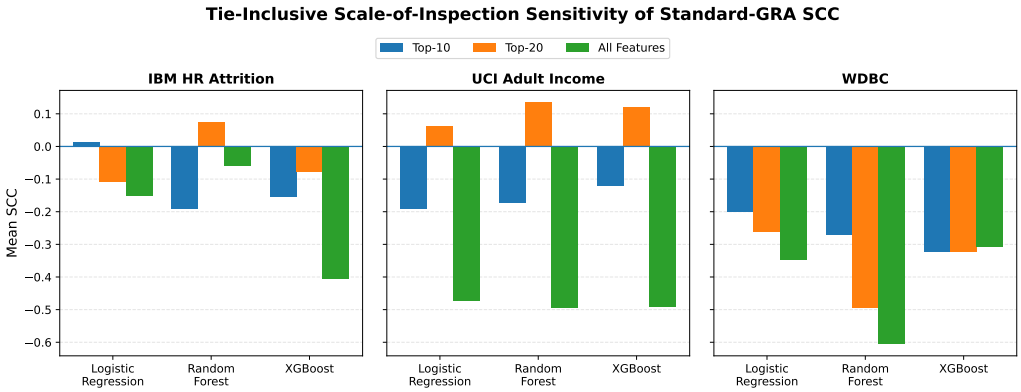



**Figure 2.** Tie-inclusive scale-of-inspection sensitivity of Standard-GRA SCC. The y-axis is the five-seed mean SCC; panels separate datasets, and bars compare nominal $K = 10$, $K = 20$, and the all-feature scope (30 features for WDBC, 52 for IBM HR Attrition, and 108 for Adult Income) across model families. No boundary tie changed the effective $K$ in these experiments.

**Table 6.** Standard-GRA SCC across inspection scopes.

| Dataset | Model | K = 10 | K = 20 | All Features |
|---|---|---|---|---|
| IBM HR Attrition | Logistic Regression | 0.013 | −0.107 | −0.151 |
| IBM HR Attrition | Random Forest | −0.191 | 0.074 | −0.060 |
| IBM HR Attrition | XGBoost | −0.156 | −0.078 | −0.406 |
| UCI Adult Income | Logistic Regression | −0.191 | 0.063 | −0.474 |
| UCI Adult Income | Random Forest | −0.173 | 0.135 | −0.495 |
| UCI Adult Income | XGBoost | −0.120 | 0.122 | −0.491 |
| WDBC | Logistic Regression | −0.200 | −0.261 | −0.347 |
| WDBC | Random Forest | −0.271 | −0.495 | −0.605 |
| WDBC | XGBoost | −0.324 | −0.322 | −0.307 |

### *4.4. GRA Operator, Distinguishing Coefficient, and Normalization Sensitivity*

Changing the GRA extrema definition from per-feature extrema to shared global extrema did not change the default SCC sign in any of the nine dataset–model combinations, but it changed magnitude to different degrees. The largest shift occurred for WDBC random forest, where SCC changed from −0.354 with the per-feature-extrema variant to −0.495 with Standard GRA (Δ = −0.141). Smaller changes were observed for Adult random forest (+0.042) and WDBC XGBoost (−0.048). Table 7 reports the Standard-versus-Legacy SCC comparison.

**Table 7.** Effect of correcting the GRA extrema definition at zeta = 0.5.

| Dataset | Model | Standard GRA | Legacy GRA | Delta (Std.-Legacy) |
|---|---|---|---|---|
| IBM HR Attrition | Logistic Regression | −0.107 | −0.116 | 0.008 |
| IBM HR Attrition | Random Forest | 0.074 | 0.059 | 0.015 |
| IBM HR Attrition | XGBoost | −0.078 | −0.088 | 0.011 |
| UCI Adult Income | Logistic Regression | 0.063 | 0.042 | 0.021 |
| UCI Adult Income | Random Forest | 0.135 | 0.093 | 0.042 |
| UCI Adult Income | XGBoost | 0.122 | 0.118 | 0.004 |
| WDBC | Logistic Regression | −0.261 | −0.223 | −0.038 |
| WDBC | Random Forest | −0.495 | −0.354 | −0.141 |
| WDBC | XGBoost | −0.322 | −0.274 | −0.048 |

Varying the distinguishing coefficient across zeta = 0.25, 0.50, and 0.75 under direct min–max normalization produced smaller changes than the major normalization changes and did not reverse the SCC sign in any dataset–model combination. The largest SCC range across the three zeta values was 0.057 for Adult XGBoost (0.105 to 0.162); other combinations varied by 0.034 or less.

Normalization had a stronger effect. The implemented robust-z-then-min–max transformation was numerically indistinguishable from direct min–max normalization, with maximum absolute differences of $2.22 \times 10^{-16}$ for IBM HR and Adult and $3.33 \times 10^{-16}$ for WDBC. This numerical result is consistent with the algebraic cancellation of a positive affine transformation followed immediately by min–max normalization; the numerical test is therefore interpreted as an implementation check rather than evidence of an additional robustification effect.

Nonlinear rank-ECDF normalization produced substantially different SCC values. At zeta = 0.5, Adult SCC increased from +0.063 to +0.406 for logistic regression, from +0.135 to +0.522 for random forest, and from +0.122 to +0.520 for XGBoost. For IBM HR random forest, rank-ECDF changed SCC from +0.074 to −0.236, reversing the sign. WDBC remained negative but became more negative for random forest (−0.495 to −0.644) and XGBoost (−0.322 to −0.411). These results show that the GRA preprocessing contract can materially alter the local–global rank relationship. Table 8 reports the normalization comparison at $\zeta = 0.5$.

**Table 8.** Standard-GRA SCC under alternative normalizations at zeta = 0.5.

| Dataset | Model | Direct Min–Max | Winsor (1%)+ Min–Max | Rank-ECDF |
|---|---|---|---|---|
| IBM HR Attrition | Logistic Regression | −0.107 | −0.128 | −0.145 |
| IBM HR Attrition | Random Forest | 0.074 | 0.048 | −0.236 |
| IBM HR Attrition | XGBoost | −0.078 | −0.055 | −0.215 |
| UCI Adult Income | Logistic Regression | 0.063 | 0.124 | 0.406 |
| UCI Adult Income | Random Forest | 0.135 | 0.173 | 0.522 |
| UCI Adult Income | XGBoost | 0.122 | 0.116 | 0.520 |
| WDBC | Logistic Regression | −0.261 | −0.272 | −0.263 |
| WDBC | Random Forest | −0.495 | −0.556 | −0.644 |
| WDBC | XGBoost | −0.322 | −0.379 | −0.411 |

### *4.5. Representation Robustness Under Disjoint Calibration*

The representation analysis used a target-free calibration subset that was disjoint from the E6 evaluation pool. The calibration/evaluation sizes were 294/1176 for IBM HR, 9768/39,074 for Adult, and 114/455 for WDBC. The high-correlation filter reduced the encoded dimensionality from 52 to 50 for IBM HR, from 108 to 105 for Adult, and from 30 to 23 for WDBC; the correlation threshold was $|\text{rho}| \geq 0.95$.

IBM HR exhibited the strongest sign change across categorical representations. Under logistic regression, SCC was −0.088 for the original 52-feature representation, +0.185 after drop-first categorical encoding, and −0.156 after high-correlation filtering. The corresponding ROC-AUC values were 0.817, 0.818, and 0.817, respectively. Adult remained close to zero but also changed sign across representations (+0.048, −0.027, and +0.032), while ROC-AUC remained essentially unchanged at approximately 0.9065.

WDBC contains no categorical predictors, so the original and drop-first representations were identical. High-correlation filtering reduced the feature count from 30 to 23 and shifted SCC from −0.162 to −0.411, whereas ROC-AUC changed only from 0.99096 to 0.99049. Within this controlled ablation, the explanation diagnostic therefore changed much more than the held-out predictive metric. This result is reported as representation sensitivity within the tested protocol, not as a universal invariance claim about predictive performance. Figure 3 visualizes these representation-dependent SCC shifts under the disjoint-calibration design. The corresponding numerical results are reported in Table 9.

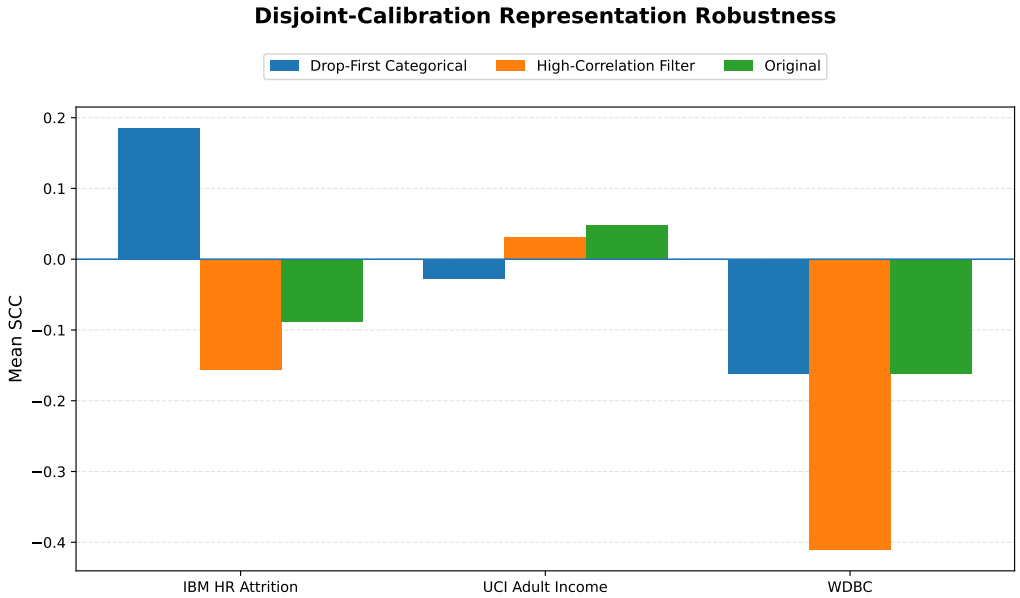



**Figure 3.** Representation robustness under disjoint calibration for logistic regression. The y-axis is the five-seed mean SCC; bars compare the Original, DropFirstCategorical, and HighCorrFilter representations for each dataset. The high-correlation filter was calibrated on a target-free subset disjoint from the evaluation pool.

**Table 9.** Logistic-regression representation robustness using disjoint calibration.

| Dataset | Representation | D | SCC | ROC-AUC | Avg. Precision |
|---|---|---|---|---|---|
| IBM HR Attrition | Original | 52 | −0.088 | 0.8174 | 0.5628 |
| IBM HR Attrition | DropFirstCategorical | 45 | 0.185 | 0.8184 | 0.5662 |
| IBM HR Attrition | HighCorrFilter | 50 | −0.156 | 0.8172 | 0.5623 |
| UCI Adult Income | Original | 108 | 0.048 | 0.9065 | 0.7646 |
| UCI Adult Income | DropFirstCategorical | 100 | −0.027 | 0.9065 | 0.7645 |
| UCI Adult Income | HighCorrFilter | 105 | 0.032 | 0.9065 | 0.7646 |
| WDBC | Original | 30 | −0.162 | 0.9910 | 0.9932 |
| WDBC | DropFirstCategorical | 30 | −0.162 | 0.9910 | 0.9932 |
| WDBC | HighCorrFilter | 23 | −0.411 | 0.9905 | 0.9930 |

### *4.6. Computational Integrity of the Frozen Journal Run*

The final code-fingerprinted run recomputed all 45 primary seed-level jobs and all 45 representation-ablation jobs from a fresh checkpoint namespace. The completion re-

ceipt records the run-configuration fingerprint, computation-code fingerprints, realized representation-filter fingerprint, cache statistics, and SHA-256 checksums for required outputs. Exact hashes are supplied in the supplementary reproducibility package rather than treated as scientific results in the main text. Historical MLMI/ICKII artifacts were hash-verified against frozen references but were not recomputed in this journal run.

Taken together, the fresh-compute results support a conditional interpretation of G-XCI: reproducibility, global-profile variation, cross-view agreement, lens choice, orientation, representation, sample scope, and inspection scale are all components of the observed result under the declared audit contract.

### *4.7. Targeted Output-Orientation and Shared-Audit-Scope Sensitivity*

The WDBC orientation audit confirmed that the sign of GRA-based cross-view agreement is conditional on the selected binary output. Under the class-1 benign orientation, mean Standard-GRA SCC was negative in all nine model-by-inspection-scale combinations. Replacing the GRA reference sequence with the complementary class-0 malignant probability reversed the mean SCC sign in all nine combinations. At K = 20, logistic regression changed from $-0.261$ to $+0.221$, random forest from $-0.495$ to $+0.307$, and XGBoost from $-0.322$ to $+0.149$. The strength of the positive pattern varied across settings; the descriptive five-seed intervals for the K = 10 logistic-regression and random-forest settings included zero, so the audit is interpreted as a mean sign-reversal result rather than universal positive agreement.

The local-attribution control isolated this change to the global relational lens rather than the absolute local Shapley ranking. Across the three WDBC model families, the maximum absolute difference between class-1 and class-0 Monte-Carlo Shapley values was at most $4.68 \times 10^{-10}$, and the minimum Spearman correlation between the corresponding aggregate absolute-Shapley rankings was 1.0. Recomputing SCC with the independently recomputed class-0 local ranking produced the same SCC as the fixed-local-rank comparison. Supplementary Table S1 and Figure S1 therefore demonstrate output-orientation dependence of the GRA reference sequence under an otherwise matched local-ranking contract.

The shared-audit-scope experiment produced a complementary result. When Standard GRA, absolute Spearman-to-output, and output permutation were all evaluated on the exact same 128 Shapley-selected held-out cases, all 27 original E3 mean-SCC signs were preserved. The maximum absolute SCC shift relative to the original protocol-specific scopes was 0.0947. For WDBC random forest, the matched-scope SCC values remained strongly separated at $-0.451$ for Standard GRA, $+0.672$ for absolute Spearman, and $+0.891$ for output permutation; IBM HR logistic regression similarly remained at $-0.106$, $+0.207$, and $+0.853$. Supplementary Table S2 and Figure S2 show that sample scope changes SCC magnitude but does not account for the observed qualitative disagreement among global audit lenses.

## 5. Discussion

The expanded study supports a more explicit audit interpretation of G-XCI. The central empirical result is not that one explanation view is consistently correct or incorrect. Rather, reproducibility of the aggregate attribution ranking, global-profile variation, cross-view ordinal agreement, proxy choice, target orientation, feature representation, audit sample scope, and inspection scope can vary or interact. This is consistent with the XAI literature that treats explanation quality as multi-dimensional and requiring explicit evaluation [4–11].

### *5.1. Explanation Stability and Cross-View Agreement Are Distinct Properties*

LAS remained relatively high in many dataset–model combinations, yet SCC could be negative, positive, or near zero. This separation is consistent with the feature-ranking

stability literature that uses repeated-run rank agreement as a same-view reproducibility construct [24,25]. In the frozen results, WDBC random forest achieved LAS = 0.931 while Standard-GRA SCC was −0.495; IBM HR random forest achieved LAS = 0.949 while SCC was only +0.074. Thus, high reproducibility of the aggregate local-attribution ranking does not imply strong cross-view agreement. OpenXAI likewise treats stability/robustness as distinct from other evaluation properties such as faithfulness [10]. G-XCI adds the separate SCC readout for conditional ordinal compatibility between declared views; it does not infer per-instance explanation reliability from LAS.

WDBC illustrates both the separation between predictive discrimination and cross-view rank agreement and the importance of output orientation. Under the retained benign-class orientation, all three model families achieved strong held-out ROC-AUC, while Standard-GRA SCC at K = 20 was negative. The targeted orientation audit then complemented the binary output probability and reversed the mean SCC sign in all nine model–scale combinations, while the aggregate absolute-Shapley ranking remained unchanged to numerical precision. Thus, the original negative sign is not an orientation-invariant property of WDBC or its predictor; it is a property of the declared local-ranking/global-reference contract.

### *5.2. The Global Audit Lens Is Part of the Result*

Standard GRA is retained as the primary instantiated lens for continuity with the originating G-XCI design and because its normalization, shared-extrema operator, and distinguishing coefficient are fully specified and auditable; this does not make it a privileged reference. E3 is therefore read as a lens-replacement experiment within a protocol that is lens-agnostic at the architectural level. Within the tested protocol, the global-lens analysis shows that SCC is not an intrinsic property of the aggregate attribution ranking. The same ranking can have a negative SCC with Standard GRA and a strongly positive SCC with another global proxy. For WDBC random forest under the retained benign-class orientation, SCC changed from −0.495 under Standard GRA to +0.680 under absolute Spearman-to-output and +0.886 under output permutation. IBM HR logistic regression similarly changed from −0.107 to +0.225 and +0.819 under those lenses.

This dependence is not a defect in SCC; it identifies what SCC actually measures. SCC is a rank agreement between two specified views, so changing the second view can legitimately change the coefficient. The targeted shared-scope audit strengthens this interpretation: after Standard GRA, absolute Spearman, and output permutation were forced onto exactly the same 128 held-out cases, all 27 original mean-SCC signs were preserved. Audit sample scope therefore affects magnitude but does not explain away the qualitative lens disagreement. Shapley explanations themselves are conditional on the feature game and dependence assumptions [12,13], and cross-paradigm studies have shown that different explanation criteria can emphasize different features [17].

The particularly high SCC values for the output-permutation proxy also require restraint. Both Monte-Carlo Shapley attribution and output permutation describe changes in the same model's output, although they aggregate perturbations differently. Their conceptual proximity makes high agreement plausible; it does not turn the permutation proxy into independent ground truth. Conversely, GRA asks a different relational question based on normalized feature–output sequence geometry. The lower or negative agreement under GRA is therefore evidence of cross-view divergence, not evidence that either method is erroneous.

### 5.3. K Is an Inspection Scale, Not a Harmless Reporting Choice

The journal experiment revises the interpretation of Adult Income. At K = 20, all three model families produced positive mean Standard-GRA SCC, but the sign was negative at K = 10 and again when the full 108-feature space was inspected. Logistic regression followed $-0.191 \rightarrow +0.063 \rightarrow -0.474$ across K = 10, 20, and all features; random forest followed $-0.173 \rightarrow +0.135 \rightarrow -0.495$; and XGBoost followed $-0.120 \rightarrow +0.122 \rightarrow -0.491$. Adult therefore exhibits pronounced inspection-scale sensitivity within the tested protocol rather than a fixed SCC sign.

We interpret K as a scale-of-inspection parameter because it determines how much of the locally ranked feature space is admitted to the cross-view comparison. This terminology is introduced operationally in the present study rather than claimed as an established convention in the literature. The underlying methodological point is straightforward: agreement among the most highly ranked features need not have the same sign or magnitude as agreement across a broader tail of predictors. The AIES comparison of feature attribution and counterfactual explanations likewise illustrates that top-ranked features can behave differently under another explanatory criterion [17], although that work does not define G-XCI's scale-of-inspection concept.

The directionality of the nominal top-$K$ construction also matters in practice. SCC($K$) asks how the declared global view orders the features first prioritized by the local attribution view; it does not test whether the two views would independently select the same top-$K$ set. The all-feature SCC should therefore be read alongside top-$K$ SCC when subset-selection asymmetry is material.

The tie-inclusive boundary rule strengthens this interpretation by ensuring that K does not become an arbitrary column-order decision when exact attribution ties occur. No boundary tie was observed in E2–E6, so the reported journal values were numerically unchanged by this safeguard; nevertheless, the rule makes the definition permutation-invariant at the selection boundary.

### 5.4. Preprocessing and Representation Belong to the Audit Contract

The representation ablation showed that different encodings of the same underlying predictors can materially alter SCC while held-out predictive behavior changes little. IBM HR logistic regression changed from SCC = $-0.088$ under the original full one-hot representation to $+0.185$ under drop-first encoding and $-0.156$ after high-correlation filtering, while ROC-AUC remained approximately 0.817–0.818. WDBC changed from $-0.162$ to $-0.411$ after high-correlation filtering under the retained benign-class orientation, whereas ROC-AUC changed from 0.99096 to 0.99049.

These findings fit an established concern in tabular XAI: correlated predictors, categorical encoding, preprocessing, and model structure can alter local explanation behavior [12,14–16]. The present study extends that concern to a multi-view reliability diagnostic. If representation changes the ranked local and global views differently, a cross-view coefficient can move substantially even when prediction is nearly unchanged. Accordingly, preprocessing is not merely a nuisance step to be hidden in implementation details; it is part of the audit contract that must be reported.

E6 therefore probes representation sensitivity under a fixed attribution construction. It does not answer how a dependence-aware or grouped-feature attribution operator would rank features under an unchanged representation; that is a distinct validation axis and remains future work rather than a like-for-like extension of E6.

E6 was intentionally isolated from the main repeated train/test pool by a target-free representation-calibration subset. This avoids fitting the high-correlation filter on E6 evaluation covariates. The main E2–E5 fixed schema, however, uses dataset-level predictor

support to align categorical vocabularies and remove global constants across seeds. This uses no target information but is not identical to a strictly train-only feature-discovery protocol. The trade-off is explicit: stable feature identities simplify rank comparison, whereas a stricter deployment simulation would estimate category support in each training split and then align heterogeneous rank spaces afterward. Future work should test that alternative rather than assuming the two contracts are equivalent.

### 5.5. Correct GRA Specification Matters, but It Does Not Create the Entire Pattern

The primary GRA operator uses shared global Δ_min and Δ_max over the complete feature-by-sample deviation matrix, consistent with published GRA formulations [21–23]. The historical per-feature-extrema implementation is retained only as a labeled ablation. Using shared extrema changed SCC magnitude most strongly for WDBC random forest ($-0.354$ to $-0.495$ under the retained benign-class orientation), but it did not reverse the default SCC sign in any of the nine dataset–model combinations.

This result has two implications. First, equation-to-code fidelity matters: a method should not be described with a global-extrema formula while executing feature-specific extrema. Second, the corrected operator does not by itself explain all observed cross-view behavior. Dataset, model, normalization, proxy, and inspection scale remain influential after the formula correction. The journal contribution is therefore not that "Standard GRA fixes G-XCI," but that explicit operator definition allows the sensitivity of the diagnostic to be audited rather than hidden.

### 5.6. The Default Robust-z → Min–Max Path Is Algebraically Redundant

E5 clarifies another methodological point inherited from the earlier pipeline. A positive affine transformation followed immediately by min–max normalization yields the same normalized sequence as direct min–max for every nonconstant sequence. The observed numerical differences of approximately $2.22 \times 10^{-16}$ to $3.33 \times 10^{-16}$ are therefore implementation-level floating-point residuals, not evidence that the preliminary robust-z stage contributes additional robustness.

This distinction is important for transparent method reporting. The default code path is retained for historical continuity, but it should not be advertised as a two-stage robust normalization. By contrast, nonlinear rank-ECDF normalization materially changed SCC, including an IBM HR random-forest sign reversal (+0.074 to $-0.236$) and substantially larger positive Adult values. Similarly, published work shows that the GRA distinguishing coefficient can affect rankings [20], which motivates the explicit $\zeta$ sensitivity analysis. In the present experiment, $\zeta$ changes were smaller than the major normalization changes and did not reverse the default sign, but that empirical pattern should not be generalized beyond the tested data.

### 5.7. Predictive Performance Does Not Substitute for Explanation Auditing

The multi-model results reinforce the separation between prediction quality and explanation agreement. On WDBC, ROC-AUC exceeded 0.987 for every model while Standard-GRA SCC remained negative for all three. In E6, representation changes produced large SCC shifts with only small ROC-AUC changes. The conclusion is not that predictive quality and explanation behavior are unrelated in general; rather, strong predictive discrimination does not, by itself, determine whether two explanation views will agree.

This separation is consistent with the motivation behind explicit explanation benchmarks and multi-metric evaluation frameworks [5,6,10]. A practical system can therefore require two distinct questions: whether the predictor performs acceptably for its task and whether its explanations are stable and compatible with the audit views selected for governance or engineering review. G-XCI addresses only the second question and should

not be interpreted as a model-selection objective unless a future study explicitly defines and validates such use.

### 5.8. Engineering Interpretation: An Explicit Audit Contract

The empirical evidence supports describing G-XCI as an audit protocol rather than a universal explanation-validity metric. A complete result should identify the predictive model; local attribution operator and baseline; selected output/class orientation; global audit lens; feature representation and preprocessing; audit sample scope; inspection scale K; and repeated-execution policy. The supplementary audits show why two of these fields cannot be treated as bookkeeping: output orientation can reverse the mean GRA-based SCC sign even when the absolute local Shapley ranking is unchanged, while matching the audit sample scope changes SCC magnitude without removing the qualitative disagreement among global lenses.

This contract-based interpretation also clarifies the appropriate response to disagreement. A negative SCC is an alert that the specified local and global rankings order the inspected features in opposing tendencies; it does not determine which ranking is preferable. Follow-up may involve checking dependence-aware attribution, grouped features, alternative global proxies, domain expertise, or data-quality issues. Such escalation is more defensible than treating disagreement as automatic evidence of model failure or explanation error.

Practical audit example. Consider a practitioner who observes strong held-out discrimination for WDBC (ROC-AUC above 0.987 for all three model families) but a negative Standard-GRA SCC at $K = 20$ ($-0.261$, $-0.495$, and $-0.322$ for logistic regression, random forest, and XGBoost, respectively). Rather than labeling the local explanation as incorrect, G-XCI prompts inspection of the declared audit contract. Complementing the retained benign-class probability with the malignant-class probability reversed the mean SCC sign at $K = 20$ to $+0.221$, $+0.307$, and $+0.149$, respectively. Across the three WDBC model families, the class-1 versus class-0 Monte-Carlo Shapley values differed by at most $4.68 \times 10^{-10}$, and the corresponding aggregate absolute-Shapley rankings had Spearman correlation 1.0. The actionable conclusion is therefore that the cross-view disagreement is output-orientation-sensitive and should be reviewed with domain context before downstream use of the GRA-based reading.

### 5.9. Limitations

Several limitations bound the present evidence. First, the evaluation includes three datasets and three model families, which broadens the controlled study but does not constitute a universal benchmark. WDBC is fully numerical, whereas IBM HR and Adult contain mixed categorical/numerical structure; additional mixed-type datasets would strengthen external validity. Second, the attribution view remains one Monte-Carlo permutation-Shapley construction with a coordinate-wise training-median baseline. For one-hot categorical blocks, that baseline can be off-manifold, so the attribution ranking is explicitly baseline-conditioned. Dependence-aware or grouped-feature alternatives were not tested.

Third, repeated seeds jointly vary the train/test split, model randomness where applicable, the 128-case Shapley audit subsample, Monte-Carlo attribution sampling, and the output-permutation randomization. LAS therefore measures end-to-end reproducibility of the aggregate ranking, not a factorial decomposition of resampling, model, and explainer variance. A design that varies these sources independently would provide a more diagnostic uncertainty analysis.

Fourth, the five-seed bootstrap intervals are deliberately descriptive. The repeated splits overlap in observations, and the five-seed realizations are not a population sample,

so the intervals are not inferential confidence intervals. Fifth, the shared-audit-scope supplement fixes the global lenses to the same 128 held-out cases, but it does not explore how conclusions change over multiple audit-set sizes or alternative deterministic sampling rules. Audit sample scope therefore remains part of the contract even though the E3 sign pattern survived the matched-scope check.

Sixth, the main fixed feature schema uses target-free dataset-level predictor support to preserve aligned feature identities across seeds. Although labels and learned numeric/model parameters remain split-specific, this contract uses covariate information beyond each individual training split. A stricter deployment-oriented protocol should estimate category vocabularies within training data only and resolve cross-seed rank-space alignment explicitly. Seventh, the output-orientation audit was performed only for the binary WDBC case. It verifies a real orientation dependence of the GRA reference sequence but does not establish how a comparable contract should be defined for multiclass outputs, logits, calibrated scores, or orientation-invariant relational operators.

Finally, the journal matrix uses a fixed Monte-Carlo budget of 32 permutations $\times$ 128 held-out Shapley-audit cases, and E6 representation robustness was evaluated with logistic regression only. These settings were chosen for computational tractability and are not claimed optimal. Higher-budget replication, additional seeds, additional model families and datasets, multi-class output-orientation audits, multiple matched audit-set sizes, dependence-aware or grouped-feature attribution, and stricter train-only representation construction remain useful extensions.

The external validity of the present evidence is therefore limited to three tabular binary-classification datasets, three model families, and one aggregate Monte-Carlo permutation-Shapley construction with a coordinate-wise median baseline. The reported results do not establish universal behavior across modalities, attribution families, multiclass or regression targets, or predictive-model families beyond those evaluated.

### *5.10. Overall Interpretation and Future Direction*

The fresh-compute findings and targeted supplementary audits support a narrower but stronger statement than a universal explanation-reliability score: G-XCI is conditional on a declared audit contract, and changing components of that contract can change the observed rank relationship. Adult Income demonstrates inspection-scale sensitivity; the global-lens analysis demonstrates proxy sensitivity that survives matched audit sample scope; the GRA ablation demonstrates operator and normalization sensitivity; the representation experiment demonstrates encoding sensitivity despite nearly unchanged predictive discrimination in several settings; and the WDBC supplement demonstrates output-orientation sensitivity while holding the absolute local attribution ranking fixed.

Future development should therefore focus less on searching for a single universal SCC threshold and more on structured contract sensitivity. Priority extensions include dependence-aware or grouped-feature attribution, additional mixed-type datasets, specialized neural tabular models, factorial separation of retraining versus explainer variability, strict train-only representation construction with rank-space alignment, multiple matched audit-set sizes, and output contracts for multiclass or otherwise non-complementary prediction spaces.

## 6. Conclusions

In the tested tabular binary-classification settings, G-XCI was evaluated as a conditional multi-view audit protocol for aggregate feature-attribution behavior rather than as a predictive model or universal explanation-correctness metric. The protocol separates three diagnostic questions: whether an aggregate ranking derived from local Monte-Carlo

Shapley attributions is reproducible across repeated executions, whether a declared global score profile varies across those executions, and whether the aggregate attribution ranking agrees ordinally with a specified global ranking within an explicit inspection scope. The implementation additionally makes the global operator, attribution baseline, output orientation, preprocessing contract, feature representation, audit sample scope, top-K policy, and computational provenance explicit.

The experiments show why that separation matters. Adult Income changes SCC sign with K, alternative global lenses can yield very different agreement for the same aggregate attribution ranking, nonlinear normalization and GRA operator choices alter SCC, and representation changes can move SCC substantially while held-out discrimination changes little. Two targeted audits sharpen the interpretation: complementing the WDBC binary output reversed the mean GRA-based SCC sign across all nine model–scale combinations while leaving the absolute local Shapley ranking unchanged, whereas forcing three global lenses onto the identical 128 held-out cases preserved every original mean-SCC sign. Output orientation and audit sample scope are therefore empirically justified components of the audit contract.

Accordingly, a complete G-XCI result should be reported together with the predictive model, attribution operator and baseline, target/output orientation, global audit lens, preprocessing/representation contract, audit sample scope, inspection scale, and repeated-execution design. Negative or low SCC is a disagreement signal that motivates further investigation, not proof that either view is wrong. Future work should add dependence-aware or grouped-feature attribution, additional mixed-type datasets and neural tabular models, stricter train-only representation construction, factorial experiments that separate resampling, model, and explainer variability, and generalized output-orientation contracts beyond binary probability complementation.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16199739/s1, Table S1: WDBC output-orientation sensitivity. "Shared-128 SCC" uses the exact Shapley-selected held-out cases. The local-rank control confirms that the aggregate absolute-Shapley ordering is unchanged by binary probability complementation; Figure S1: WDBC Standard-GRA SCC under benign (class 1) and malignant (class 0) output orientations across model families and inspection scopes; Table S2: Shared-audit-scope comparison against the original E3 scopes. All three global lenses use the exact 128 Shapley-selected held-out cases in the shared-scope column; Figure S2: Cross-view SCC under the matched 128-case audit scope for Standard GRA, absolute Spearman-to-output, and output permutation; Table S3: Individual seed-level E2 results for the primary Standard-GRA analysis at nominal $K = 20$. Predictive metrics are held-out controls and are not treated as evidence of explanation correctness; the exact frozen IBM HR and Adult input files and their hash-verified parent source archive; the authoritative v3.1.2 journal notebook, supplementary audit notebook, and verified publication-figure generation notebook; the machine-readable E0–E6 main-experiment outputs; the S1–S2 seed-level and summary audit outputs; the machine-readable Table S3 source; run manifests; software-environment information; and completion/checksum receipts.

**Author Contributions:** Conceptualization, C.-H.L. and W.-C.Y.; methodology, C.-H.L.; software, C.-H.L.; validation, C.-H.L. and Y.L.; formal analysis, C.-H.L.; investigation, C.-H.L. and Y.L.; resources, Y.L.; writing—original draft preparation, C.-H.L.; writing—review and editing, C.-H.L. and Y.L.; administration, Y.-S.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable. This study performed secondary analysis of publicly available benchmark datasets and did not recruit participants, conduct interventions, or

access private identifiable information. The IBM HR benchmark is a fictional/synthetic dataset; the Adult Income and WDBC data were accessed through public benchmark repositories.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The Adult Income dataset is publicly available from the UCI Machine Learning Repository [28]. The IBM HR Analytics Employee Attrition & Performance dataset is publicly available through Kaggle [29]. The Wisconsin Diagnostic Breast Cancer dataset is publicly available from the UCI Machine Learning Repository [30] and was loaded in the journal notebook through scikit-learn. Supplementary File S1 includes the exact frozen IBM HR and Adult input files used by the completed v3.1.2 journal computation together with their parent frozen source archive. The supplied parent archive has SHA-256 9e92781e9f2823ebdde6c467150f2f4bd49e310e19dea26108c1ce8f613bf100, exactly matching the value recorded in the authoritative completed-run manifest. Supplementary File S1 also includes the authoritative journal notebook, the supplementary audit notebook, machine-readable E0–E6 main-experiment tables, targeted S1–S2 audit outputs, the machine-readable S3 seed-level table, run manifests, software-environment information, and completion/checksum receipts. No proprietary data were used in the reported experiments.

**Acknowledgments:** During the preparation of this manuscript and study, the authors used ChatGPT (OpenAI), Claude (Anthropic), Google Gemini, Microsoft Copilot, SciSpace, and Perplexity Pro for the limited purposes described in Section 3.14. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this publication.

**Conflicts of Interest:** Author Wen-Chun Yang is employed by the company Edrington Taiwan Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| G-XCI | Grey–Shapley Cross-Consistency Index |
| XAI | Explainable Artificial Intelligence |
| LAS | Local Attribution Stability |
| GRS | Global Relational Stability |
| SCC | Structural Consistency Coefficient |
| GRA | Grey Relational Analysis |
| SHAP | SHapley Additive exPlanations |
| WDBC | Wisconsin Diagnostic Breast Cancer |
| ROC-AUC | Area Under the Receiver Operating Characteristic Curve |

## References

1. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
2. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
3. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume PMLR 70, pp. 3319–3328.
4. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems 31*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; pp. 9505–9515.
5. Hooker, S.; Erhan, D.; Kindermans, P.-J.; Kim, B. A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 9737–9748.
6. Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D.; Ravikumar, P. On the (In)fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 10965–10976.

7. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–9 February 2020; pp. 180–186. [CrossRef]
8. Slack, D.; Hilgard, A.; Singh, S.; Lakkaraju, H. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In *Advances in Neural Information Processing Systems 34*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; pp. 9391–9404.
9. Zhou, Z.; Hooker, G.; Wang, F. S-LIME: Stabilized-LIME for Model Explanation. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual Event, 14–18 August 2021; pp. 2429–2438.. [CrossRef]
10. Agarwal, C.; Krishna, S.; Saxena, E.; Pawelczyk, M.; Johnson, N.; Puri, I.; Zitnik, M.; Lakkaraju, H. OpenXAI: Towards a Transparent Evaluation of Model Explanations. In *Advances in Neural Information Processing Systems 35, Datasets and Benchmarks Track*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2022; pp. 15784–15799.. [CrossRef]
11. Kelodjou, G.; Rozé, L.; Masson, V.; Galárraga, L.; Gaudel, R.; Tchuente, M.; Termier, A. Shaping up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 13094–13103. [CrossRef]
12. Aas, K.; Jullum, M.; Løland, A. Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *Artif. Intell.* **2021**, *298*, 103502. [CrossRef]
13. Kumar, I.E.; Venkatasubramanian, S.; Scheidegger, C.; Friedler, S. Problems with Shapley-Value-Based Explanations as Feature Importance Measures. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; PMLR 119, pp. 5491–5500.
14. Rahnama, A.H.A.; Bütepage, J.; Geurts, P.; Boström, H. Can Local Explanation Techniques Explain Linear Additive Models? *Data Min. Knowl. Discov.* **2024**, *38*, 237–280. [CrossRef]
15. Elkhawaga, G.; Abuelkheir, M.; Reichert, M. Explainability of Predictive Process Monitoring Results: Can You See My Data Issues? *Appl. Sci.* **2022**, *12*, 8192. [CrossRef]
16. El-Khawaga, G.; Abu-Elkheir, M.; Reichert, M. XAI in the Context of Predictive Process Monitoring: An Empirical Analysis Framework. *Algorithms* **2022**, *15*, 199. [CrossRef]
17. Mothilal, R.K.; Mahajan, D.; Tan, C.; Sharma, A. Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, 19–21 May 2021. [CrossRef]
18. Deng, J.-L. Control Problems of Grey Systems. *Syst. Control Lett.* **1982**, *1*, 288–294. [CrossRef]
19. Liu, S.; Yang, Y.; Xie, N.; Forrest, J. New Progress of Grey System Theory in the New Millennium. *Grey Syst. Theory Appl.* **2016**, *6*, 2–31. [CrossRef]
20. Mahmoudi, A.; Javed, S.A.; Liu, S.; Deng, X. Distinguishing Coefficient Driven Sensitivity Analysis of GRA Model for Intelligent Decisions: Application in Project Management. *Technol. Econ. Dev. Econ.* **2020**, *26*, 621–641. [CrossRef]
21. Pitchipoo, P.; Venkumar, P.; Rajakarunakaran, S. Modeling and Development of a Decision Support System for Supplier Selection in the Process Industry. *J. Ind. Eng. Int.* **2013**, *9*, 23. [CrossRef]
22. Shan, Z.; Wu, W.; Lei, Y.; Zhao, B. A New Fuzzy Rule Based Multi-Objective Optimization Method for Cross-Scale Injection Molding of Protein Electrophoresis Microfluidic Chips. *Sci. Rep.* **2022**, *12*, 13159. [CrossRef] [PubMed]
23. Liu, S.; Yang, Y.; Forrest, J. *Grey Data Analysis: Methods, Models and Applications*; Springer: Singapore, 2017. [CrossRef]
24. Wang, H.; Yang, F.; Luo, Z. An Experimental Study of the Intrinsic Stability of Random Forest Variable Importance Measures. *BMC Bioinform.* **2016**, *17*, 60. [CrossRef] [PubMed]
25. Nogueira, S.; Sechidis, K.; Brown, G. On the Use of Spearman's Rho to Measure the Stability of Feature Rankings. In *Pattern Recognition and Image Analysis*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; pp. 381–391. [CrossRef]
26. Kendall, M.G. The Treatment of Ties in Ranking Problems. *Biometrika* **1945**, *33*, 239–251. [CrossRef] [PubMed]
27. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef] [PubMed]
28. Becker, B.; Kohavi, R. Adult [Dataset]. UCI Machine Learning Repository, 1996. Available online: https://archive.ics.uci.edu/dataset/2/adult (accessed on 27 August 2026). [CrossRef]
29. IBM HR Analytics Employee Attrition & Performance [Dataset]. Kaggle. Available online: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset (accessed on 27 August 2026).
30. Wolberg, W.; Mangasarian, O.; Street, N.; Street, W. Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. 1993. Available online: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic (accessed on 27 August 2026). [CrossRef]
31. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.