MULTI-LEVEL ADAPTATION OF ROUGH INCLUSION CLASSIFIERS WITH FEATURE AND PAIRWISE CORRECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

How much structure should be added to a rough inclusion classifier when its original similarity score is insufficient? We study a multi-level adaptation procedure that uses training validation to select among the unmodified classifier, score calibration, univariate value-dependent corrections, and constrained pairwise correction tables. The complete base-tuning procedure is repeated within cross-fitting to learn and evaluate these extensions. Across 240 train/test splits of eight binary datasets, macro-averaged balanced accuracy increases from 82.05% to 85.93%. Almost all of this improvement comes from calibration and univariate effects: allowing pairwise corrections adds only 0.02 percentage points over the selected univariate control. Although the pairwise mechanism reliably recovers strong synthetic XOR interactions, validation gains only weakly predict test gains on real datasets. An EBM correction layer initialized with the same base scores remains competitive, limiting claims of a distinctive performance advantage. These findings separate the ability to represent interactions from the ability to select interactions that generalize. Our contribution is an auditable adaptation of this classifier family and an empirical analysis of when additional correction levels help. Within the studied setting, the results identify reliable structure selection, rather than additional pairwise representational capacity, as the more pressing challenge.

1 INTRODUCTION

Similarity-based classification can miss information carried by combinations of feature values. In balanced XOR, either feature alone is uninformative, while their joint configuration determines the label. For rough inclusion classifiers, should a prediction therefore be adjusted globally, by individual feature values, or by selected combinations?

We investigate this question as a continuation of sinusoidal weighting for rough inclusion classification (Artiemjew & Samojluk, 2026; Samojluk & Artiemjew, 2026). The starting point is the frozen v3.1.1 implementation. We add a correction layer without changing its similarity calculations. Calibration adjusts the scale and threshold of its score. Univariate tables then assign corrections to value ranges of each feature. Finally, pairwise tables describe additional corrections associated with two conditions holding together. Positive entries favor class 1 and negative entries favor class 0. A gate, when used, controls these pair contributions; it is not a separate source of rewards.

Training validation selects among the original base, calibration, univariate correction, and pairwise correction, including the option to reject an extension. A three-dimensional view of a pair table places its two inputs on the horizontal axes and the correction on the vertical axis. It visualizes a bivariate function without adding a measured input dimension.

Our contributions are threefold. First, we specify an executable correction construction for this rough inclusion family, including constrained pair tables, nested base tuning, and structure selection. Second, we separate the sources of observed improvement across 240 archived splits, showing that lower levels explain nearly all of the average gain. Third, we add matched-base EBM, duplicate-grouped, and continuous-interaction controls, together with a support-gate ablation. These experiments identify both useful behavior and unresolved weaknesses. They do not establish state-of-the-art performance or the first use of feature interactions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Throughout, *v4.0* denotes the frozen algorithm and *v4.5* its support-gate laboratory variant. The manuscript version is v5. Historical and new measurements are reported separately.

2 Related Work

Additive models with interactions. GA^{2}M combines main effects with selected pairwise functions and introduces FAST interaction ranking (Lou et al., 2013). EBM provides an interpretable boosting implementation of this closely related model family (Nori et al., 2019). GAMI-Net uses a staged main-effect and interaction architecture with sparsity, heredity, and clarity constraints (Yang et al., 2021). GAMI-Tree incorporates pair filtering and purification into model-based boosting (Hu et al., 2023), while NODE-GAM includes a differentiable pairwise extension (Chang et al., 2022). Thus, neither the sum of univariate and bivariate functions nor its surface visualization is new. Our correction tables use a different constrained optimization procedure and retain an externally tuned rough inclusion score.

Correction of an existing predictor. Learning a second level from out-of-fold predictions has a precedent in stacked generalization (Ting & Witten, 1999). More directly, the EBM API accepts an external `init_score` in both fitting and prediction (InterpretML, 2026); we evaluate this control rather than relying only on standalone EBM. Cyclic Boosting also learns contributions for discretized feature groups (Wick et al., 2019). Its implementation supports prior predictions and an optional two-dimensional neutralization smoother (Blue Yonder, 2026a;b). It is therefore a close alternative at the level of correction tables, not merely a generic boosting baseline.

Separating and selecting interactions. Functional ANOVA purification formalizes the separation of lower- and higher-order effects relative to a distribution (Lengerich et al., 2020). Our constraints apply only to each table's empirical axes and do not guarantee global identifiability. A table cell can also be represented as a weighted conjunction of conditions, connecting the construction to rule ensembles (Friedman & Popescu, 2008). Hierarchical interaction models impose different structural restrictions (Bien et al., 2013); we do not require a pair to have a predictive parent main effect. Recent analysis of pure interactions in tree methods further motivates testing signals that lack marginal effects (Blum et al., 2025). Earlier rough-set work already used rules to identify interaction candidates for a logit model (Ong et al., 2004).

These precedents substantially limit a broad novelty claim. The object studied here is the particular integration with a rough inclusion base, support-constrained fitting, and optional model complexity. Its scientific value must be judged through controlled comparisons. Supplement J maps the closest mechanisms and the scope of source verification; the review is targeted, not an exhaustive priority search.

3 Multi-Level Correction Model

3.1 Base score and univariate correction

For binary labels $y \in \{0, 1\}$, the base computes a mean comparison cost $C_c(x)$ to each class. Lower cost indicates greater similarity. Define

$$r(x) = \log \max(C_0(x), 10^{-15}) - \log \max(C_1(x), 10^{-15}), \qquad z(x) = b + a\frac{r(x) - \mu}{\sigma}, \quad a \geq 0. \tag{1}$$

The calibration parameters are learned with class-balanced logistic loss on cross-fitted base scores. They remain fixed during subsequent correction fitting. The base includes 283 configurations, with normalized absolute differences and optional sinusoidal modulation. In the audited implementation, that modulation depends on the *training row index*, not the feature index. It can therefore be sensitive to training order. Supplement A specifies the exact costs and their relationship to earlier descriptions.

For each feature, a training-only binning map b_j defines a piecewise-constant function f_j. We learn a correction for each value range or category, not one global feature weight. Centering the bin encoding makes its empirical mean zero; a separate intercept absorbs the overall shift. Univariate corrections can adjust a decision even when there is no interaction.

3.2 Pairwise tables and support constraints

Let $\mathcal{P}$ contain the selected feature pairs. A corrected score has the form

$$s(x) = z(x) + \beta_0 + \sum_j f_j(x_j) + g(z(x)) \sum_{(j,k)\in\mathcal{P}} H_{jk}[b_j^{jk}(x_j), b_k^{jk}(x_k)]. \tag{2}$$

The pair axes may differ from the univariate axes. The ungated family has $g = 1$; the original optional gate is $g_{\text{conf}}(z) = [1 + (z/2)^2]^{-1}$. The decision is class 1 iff $s(x) > 0$, with ties assigned to class 0. Dry-base and calibration-only candidates omit the later terms.

For one table, let N_{uv} be the number of training records in cell (u, v). Unsupported cells, with $N_{uv} < 3$, have zero correction. Supported entries satisfy

$$\sum_v N_{uv}H_{uv} = 0 \quad \text{for every row } u, \qquad \sum_u N_{uv}H_{uv} = 0 \quad \text{for every column } v. \tag{3}$$

We optimize in a null-space basis of these constraints. This prevents a table from freely reproducing effects of its own row or column categories. It is a local empirical separation: different axis resolutions, dependent features, and the base-dependent confidence gate prevent interpreting the whole model as a globally purified two-way ANOVA.

The original table penalty is $\frac{\alpha}{2}\sum_{N_{uv}\geq 3}(1 + 5/N_{uv})H_{uv}^2$, alongside ridge penalties for univariate terms. This penalizes corrections in less populated cells more strongly. After screening, all retained univariate and pair coefficients are fitted jointly by class-balanced penalized logistic optimization with fixed offset z.

What XOR illustrates. With equal counts in a 2×2 table, the pattern $\left[\begin{smallmatrix} -h & h \\ h & -h \end{smallmatrix}\right]$ satisfies Equation 3. Neither feature alone distinguishes the classes, but opposite conditions receive a positive correction. XOR is one representable pattern, not the definition of the entire pair mechanism. In the earlier toy example, the base, matched univariate model, and pair model classify 20/36, 18/36, and 36/36 test points correctly. Its learned calibration suppresses the base score; the pair table supplies the decision. Supplement G gives the fitted numbers and a threshold-gap counterexample.

3.3 Support-based gate ablation

The confidence gate may attenuate a correction when the base is confidently wrong. Laboratory v4.5 replaces that candidate family by a cell-support gate $g_{uv} = N_{uv}/(N_{uv} + 5)$ for supported cells, and zero otherwise. We constrain the *effective* table H in Equation 3, define a latent table $M_{uv} = H_{uv}/g_{uv}$, and penalize M. The resulting penalty is

$$\frac{\alpha}{2}\sum_{N_{uv}\geq 3}(1 + 5/N_{uv})\frac{H_{uv}^2}{g_{uv}^2} = \frac{\alpha}{2}\sum_{N_{uv}\geq 3}(1 + 5/N_{uv})^3 H_{uv}^2. \tag{4}$$

Prediction adds H once, using $g = 1$ in Equation 2. Relative to ungated tables, this changes regularization rather than the feasible effective-table space. Multiplying an already fitted table by unequal cell gates would generally break its zero margins. Both Auto variants retain ungated pairs and the same total candidate count. This ablation leaves the minimum support and binning restrictions unchanged.

4 Learning and Structure Selection

The procedure uses only outer-training data for all tuning. Five head-validation folds evaluate the correction recipes. In each fold, the base is retuned on that fold's training subset; three further cross-fitting folds, each with a separately retuned base, provide scores for learning the correction. The corresponding validation subset is scored by the base fitted on the full head-training subset. Final corrections use the five-fold out-of-fold vector from the head-validation stage, and the final base is refitted on all outer-training observations. Preprocessing is fitted within the corresponding training partition. Supplement B gives the full sequence and code mapping.

The base bank is selected by $J = \overline{\text{BA}} - 0.1\ \text{SD}_{\text{pop}}(\text{BA})$ over five folds. Each Auto version considers 86 recipes: one dry base, one calibrated base, 12 univariate, 36 ungated pair, and 36 gated pair

recipes. Binning variants, four regularization strengths, and pair budgets of one, three, or five define the bank. Training-only screening does not require marginally informative features, and a residual quadratic-gain criterion ranks supported pairs before joint fitting.

A more complex recipe must improve mean validation BA by more than both 0.2 percentage points and the paired standard error, improve at least three of five folds, and improve J. Pair candidates are compared with the selected control allowing at most univariate correction. This is a conservative heuristic, not a calibrated significance test. "Calibration" and "1D" in the experiments denote selected stages that can fall back to a simpler model; they are not forced component fits. Test accuracy can decrease at a later stage despite satisfying the training rule.

5 Experimental Setup

We distinguish two real-data cohorts. The historical cohort comprises eight binary datasets—Australian, Diabetes, Cleveland, Wisconsin Original, Hepatitis, Ionosphere, Banknote, and Sonar—with 30 stratified 70/30 splits each (seeds 4000–4029). The new laboratory cohort uses five fresh splits of each of the same datasets (seeds 4500–4504). Fresh seeds do not make reused datasets an independent confirmation. Sample sizes range from 155 to 1,372 and feature counts from 4 to 60. Supplement D specifies preprocessing and label mappings.

Our primary metric is balanced accuracy, $\mathrm{BA} = (\mathrm{TPR} + \mathrm{TNR})/2$. Macro means weight datasets equally; differences are percentage points (pp). Across-split standard deviations are descriptive. Test sets overlap, and we do not treat the 240 splits as independent datasets or use them to justify a universal significance claim.

The new cohort compares Auto 4.0, Auto 4.5, standalone EBM, and *base + EBM*. The latter receives the same calibrated cross-fitted base scores through `init_score` during fitting and the calibrated refitted-base score at prediction. The base is added exactly once. Both EBM controls use InterpretML 0.7.8, up to five selected interactions, and the same six configurations and five head-validation folds. SVM–RBF uses 12 configurations with fold-local median imputation and standardization. Search spaces and computational budgets differ; these are protocol-specific comparisons. Supplement E gives exact settings and the scope of EBM's internal early stopping.

We additionally group identical feature patterns on Wisconsin, Banknote, and Ionosphere, keeping groups intact in the outer split, every inner fold, and EBM bags. Wisconsin requires a stronger grouping because distinct rows with missing entries can become identical after median imputation: we use connected components of compatible observed patterns. The main grouped panel contains five splits per dataset with no train/test collisions either before or after imputation (Supplement F). Finally, 75 synthetic fits vary interaction strength and sample size; five further fits test rare exceptions to a strong base signal.

6 Results and Ablations

6.1 Sources of improvement and selected structure

Historical macro BA rises from 82.05% for the dry base to 85.93% for Auto 4.0, a gain of 3.88 pp. The up-to-univariate control already reaches 85.91%; the additional Auto gain is only 0.016 pp, rounded to 0.02 pp. Figure 1 shows where improvement occurs. Wisconsin benefits mainly from calibration (92.95% to 97.22% at that stage), Banknote from univariate effects (87.86% to 99.22%), and Ionosphere gains another 0.95 pp when pairs are allowed (89.72% to 90.68%). Auto improves over the base on five of eight datasets and decreases on the other three.

Auto 4.0 has higher historical mean BA than the tuned SVM on Australian, Diabetes, Wisconsin, and Hepatitis. The margins are 0.47, 0.75, 0.34, and 1.59 pp, respectively. These are useful dataset-specific successes. However, the base already exceeds SVM on Australian and Hepatitis, and Auto decreases its score there. Only Diabetes and Wisconsin move from below this SVM baseline to above it after adaptation. Overall SVM averages 87.00%, exceeding Auto's 85.93%; full per-dataset variation is reported in Supplement D.

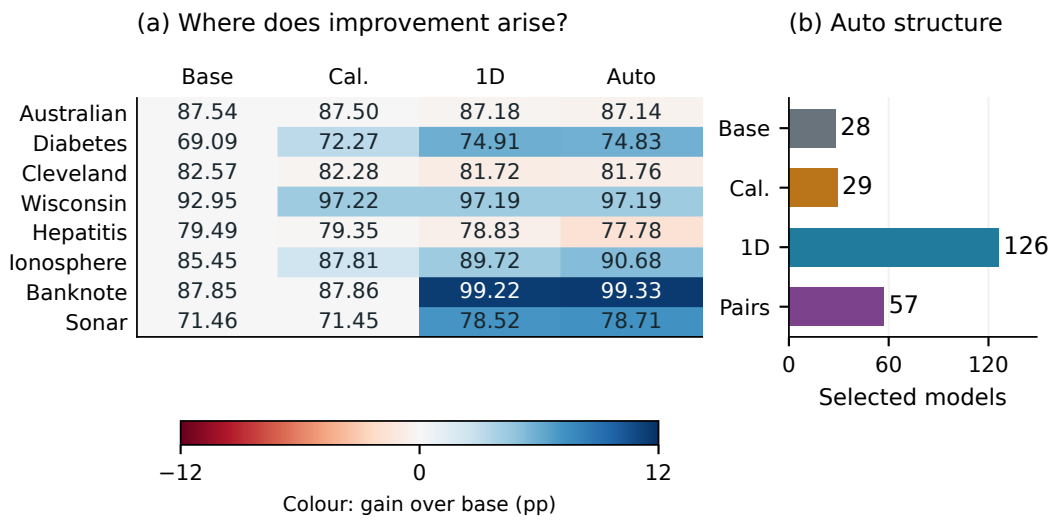

Figure 1: Historical v4.0, eight datasets × 30 splits. (a) Numbers are mean test BA (%); colour encodes the difference from each dataset's base. Cal. and 1D are selected stages allowing fallback. (b) Auto selects different levels: 28 base, 29 calibration, 126 univariate, and 57 pair models (27 ungated, 30 confidence-gated). Structure adaptation does not itself establish reliable selection of test improvements.

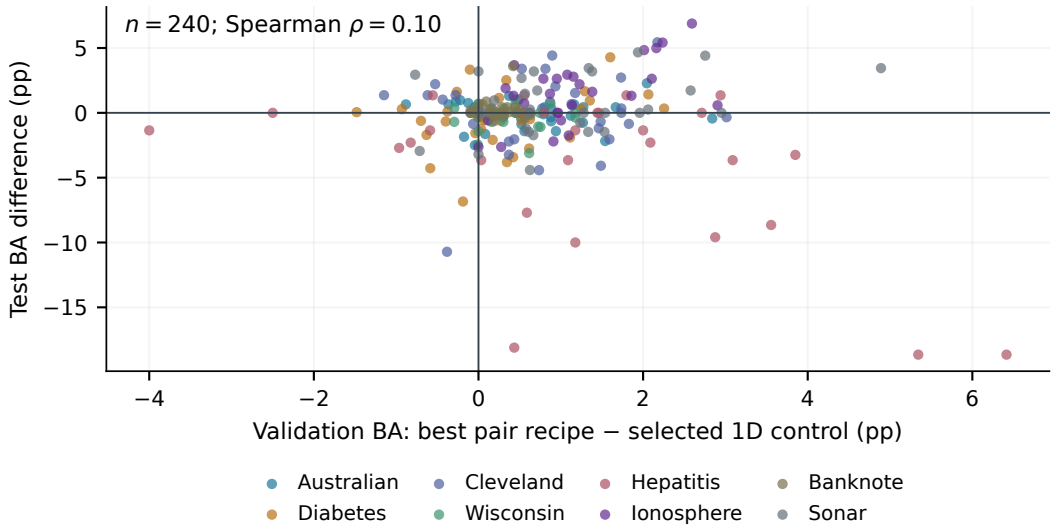

Figure 2: All 240 historical splits. Each point compares the best pair recipe by training-CV objective, before fallback, with the selected up-to-univariate control. The same recipe is evaluated on the test set. Zero lines distinguish gains and losses; rejected candidates are not replaced by artificial $(0, 0)$ points. Actual Auto decisions are evaluated separately: among 57 pair choices, 26 improve test BA, 11 tie, and 20 decrease it.

6.2 Do validation gains predict useful pairs?

Figure 2 does not provide strong evidence that the selector consistently anticipates out-of-sample pair benefits. Of 181 candidates with positive mean validation gain, only 77 also improve test BA. Pooled Spearman correlation is 0.10 and falls to 0.04 after centering gains within datasets. Ionosphere is more encouraging ($\rho = 0.56$), but behavior varies across datasets. Among the 57 actual Auto pair choices, mean test gain over the selected univariate control is just 0.066 pp. The candidate plotted

Table 1: New ordinary splits: mean test BA (%), five splits per dataset. All methods use the same outer partitions. EBM = standalone EBM; Base+EBM uses calibrated v3.1.1 via `init_score`. Budgets differ. Complete means and standard deviations are in Supplement E.

Dataset	Auto 4.0	Auto 4.5	Base+EBM	EBM	SVM
Australian	86.54	86.57	88.22	88.11	86.72
Diabetes	73.23	72.79	72.14	73.71	71.77
Cleveland	82.69	82.14	82.89	84.22	83.84
Wisconsin	95.84	95.84	96.19	96.75	96.60
Hepatitis	73.41	75.03	72.08	71.78	79.00
Ionosphere	92.09	92.09	91.83	90.95	93.73
Banknote	99.26	99.26	100.00	99.96	100.00
Sonar	80.00	79.60	83.63	81.27	86.77
Macro mean	85.38	85.41	85.87	85.84	87.30

Table 2: Grouped-pattern test BA, mean (SD) in %, five splits each. Wisconsin uses imputation-compatible groups; the other datasets use exact feature patterns. All 15 main tests have zero train/test pattern overlap before and after imputation.

Dataset	Auto 4.5	Base+EBM	SVM
Wisconsin	96.74 (1.80)	96.39 (1.45)	96.81 (1.14)
Ionosphere	88.69 (4.22)	92.82 (2.20)	94.53 (2.71)
Banknote	98.74 (1.00)	99.74 (0.47)	99.78 (0.49)

need not be the eligible recipe ultimately selected by Auto. Supplement H includes dataset panels and a matched-configuration sensitivity analysis.

6.3 Matched-base EBM and duplicate-grouped controls

The matched-base control is competitive with both correction variants (Table 1). Macro BA is 85.38% for Auto 4.0, 85.41% for Auto 4.5, 85.87% for base+EBM, 85.84% for standalone EBM, and 87.30% for SVM. The selected univariate control averages 85.58%. Thus, the new gate adds only 0.033 pp over Auto 4.0 and does not establish an aggregate advantage over a known correction method. Dataset-specific advantages remain, but the small five-split cohort cannot support a general superiority claim.

Table 2 tests generalization to held-out pattern groups. Auto remains competitive on Wisconsin, while base+EBM and SVM are stronger on Banknote and Ionosphere. Grouping changes the training and test composition, so differences from ordinary splits are not causal estimates of a "duplicate effect." This control addresses exact or imputation-induced overlap, not subject-level dependence or all forms of dataset leakage.

6.4 A continuum from main effects to XOR

Generate eight independent $\mathrm{Uniform}(-1, 1)$ features. For the first two, let $q(x) = 1$ when their signs differ and -1 otherwise. Define

$$\Pr(Y = 1 \mid x) = \mathrm{sigmoid}\left(3\frac{(1-\lambda)\sqrt{3}\,x_3 + \lambda q(x)}{\sqrt{(1-\lambda)^2 + \lambda^2}}\right). \tag{5}$$

We vary $\lambda \in \{0, .25, .5, .75, 1\}$ and training size $n \in \{160, 400, 800\}$, with five replicates and 4,000 independently generated test records each. The true pair competes among 28 feature pairs; it is not supplied to the learner. Latent-score variance is constant, although Bayes accuracy changes. Common random numbers couple strengths and nested training sizes within a replicate (Supplement G).

At $\lambda \geq 0.5$, Auto 4.5 includes the true pair in all 45 fits. For $n = 800, \lambda = 1$, it reaches 93.51% BA, 43.09 pp above the selected univariate control. At $\lambda = 0$, it selects unnecessary pairs in 3/15 fits. The transition is therefore informative but imperfect (Figure 3). Auto 4.0 also performs well on strong

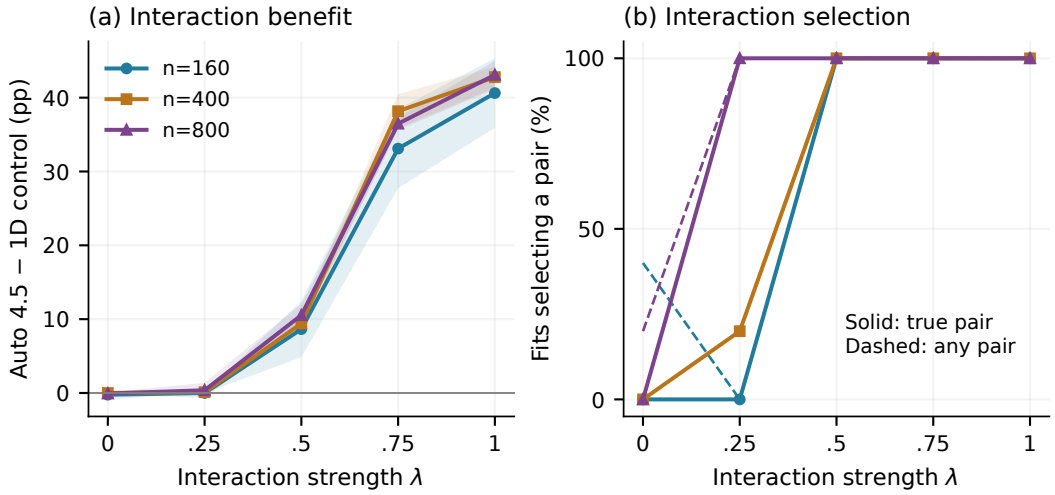

Figure 3: Noisy interaction continuum, Auto 4.5. (a) Mean BA gain over the selected univariate control, with bands of one across-replicate SD. (b) Frequency of selecting any pair (dashed) and the true pair (solid). Each point summarizes five fits. Strong-interaction recovery coexists with occasional unnecessary pair selection at $\lambda = 0$.

XOR: this demonstrates the pair machinery, not a unique benefit of the new gate. These synthetic runs do not rank our method against EBM or all alternative interaction learners.

The rare-exception stress test is less favorable. Auto 4.0 and 4.5 make identical predictions, averaging 92.01% overall BA but only 21.12% accuracy on records where the local rule contradicts the base signal. This shows why overall accuracy alone can hide failure to recover an important small subgroup. The new gate does not by itself solve that problem (Supplement I).

7 Limitations and Discussion

The construction adapts structure, but real-data selection of useful pairs remains weak. The support gate is a regularization change; empirical zero margins are not a global uniqueness theorem. The base's row-order dependence and training-set-dependent prediction cost remain. A separate single-thread runtime pilot found median tuned fit times of 13.03 s for Auto 4.0, 0.55 s for SVM, and 27.39 s for EBM with pairs, but EBM had much faster batched prediction (6.25 versus 161.38 microseconds per record). These medians use unequal searches and a small fixed-seed pilot, not the new laboratory timings (Supplement K).

The datasets are small and reused during development; grouping does not remove every dependency. No multiclass, deployment, or broad state-of-the-art claim is supported. The most defensible distinction is the audited integration with this base family and explicit accounting of model complexity and gain. Further gains require better selection, support modeling, or a demonstrated cost advantage, rather than a stronger metaphor about dimensionality.

8 Conclusion

We presented a multi-level correction procedure for binary rough inclusion classification. Its principal real-data benefit comes from calibration and univariate effects, while pair tables recover strong synthetic interactions and yield a modest dataset-specific benefit on Ionosphere. The algorithm selects different structures, but selection frequency is not equivalent to generalization success. Matched-base EBM and grouped-pattern controls narrow the claim to a reproducible adaptation of this classifier family with identifiable strengths and limitations.

Reproducibility Statement

The accompanying source package regenerates the figures and tables from preserved measurements. It includes the correction code, data definitions, protocols, numeric results, and audit records. The full laboratory archive additionally contains the 140 fitted models and 240 historical checkpoints. The separate supplement specifies the base classifier, nested fitting, comparator budgets, grouping rules, generators, and code mapping. No models were retuned when preparing this manuscript revision.

AI Use Statement

ChatGPT/Codex assisted with code inspection and implementation, experimental scripts and execution, literature retrieval, manuscript drafting and translation, and figure and table generation. Numerical claims are linked to saved results and automated replay checks. These checks do not replace the authors' scientific assessment; this consultation draft requires author review before submission.

Ethics Statement

The experiments use public benchmark datasets and synthetic data. No new personal data were collected. The results do not establish clinical, financial, or other operational suitability, and correction tables are predictive rather than causal explanations.

References

Piotr Artiemjew and Artur Samojluk. Dynamic weighting with sinusoidal modulation for rough inclusion classifiers. In *Proceedings of ICAART 2026*, volume 3, pp. 2267–2272. SCITEPRESS, 2026. doi: 10.5220/0014218100004052.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013. doi: 10.1214/13-AOS1096. URL https://arxiv.org/abs/1205.5050.

Blue Yonder. Cyclic boosting: Concepts and api reference, 2026a. URL https://cyclic-boosting.readthedocs.io/en/latest/cyclic_boosting.html. Accessed 16 September 2026; documentation of the current implementation.

Blue Yonder. Cyclic boosting: Source code for smoothing.multidim, 2026b. URL https://cyclic-boosting.readthedocs.io/en/latest/_modules/cyclic_boosting/smoothing/multidim.html. Accessed 16 September 2026; Neutralize2DMetaSmoother.

Ricardo Blum, Munir Hiabu, Enno Mammen, and Joseph T. Meyer. Pure interaction effects unseen by random forests. *Computational Statistics & Data Analysis*, 212:108237, 2025. doi: 10.1016/j.csda.2025.108237. URL https://arxiv.org/abs/2406.15500.

Chun-Hao Chang, Rich Caruana, and Anna Goldenberg. NODE-GAM: Neural generalized additive model for interpretable deep learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.01613.

Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008. doi: 10.1214/07-AOAS148. URL https://arxiv.org/abs/0811.1679.

Linwei Hu, Jie Chen, and Vijayan N. Nair. Using model-based trees with boosting to fit low-order functional ANOVA models. arXiv:2207.06950, version 5, 2023. URL https://arxiv.org/abs/2207.06950.

InterpretML. ExplainableBoostingClassifier: Api reference, 2026. URL https://interpret.ml/docs/python/api/ExplainableBoostingClassifier.html. Accessed 16 September 2026.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *Proceedings of AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2402–2412, 2020. URL https://proceedings.mlr.press/v108/lengerich20a.html.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of ACM SIGKDD*, pp. 623–631, 2013. doi: 10.1145/2487575.2487579. URL https://www.cs.cornell.edu/~yinlou/papers/lou-kdd13.pdf.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A unified framework for machine learning interpretability. arXiv:1909.09223, 2019. URL https://arxiv.org/abs/1909.09223.

Chorng-Shyong Ong, Jih-Jeng Huang, and Gwo-Hshiung Tzeng. Using rough set theory for detecting the interaction terms in a generalized logit model. In *Rough Sets and Current Trends in Computing (RSCTC 2004)*, volume 3066 of *Lecture Notes in Computer Science*, pp. 624–629. Springer, 2004. doi: 10.1007/978-3-540-25929-9_77.

Artur Samojluk and Piotr Artiemjew. On the dual nature of sinusoidal rough inclusion classifiers: From structural immunity to parametric optimization. ISD 2026, conference programme, 2026. URL https://easychair.org/smart-program/ISD2026/2026-09-02.html. DOI unconfirmed as of 17 September 2026.

Kai Ming Ting and Ian H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999. URL https://arxiv.org/abs/1105.5466.

Felix Wick, Ulrich Kerzel, and Michael Feindt. Cyclic boosting – an explainable supervised machine learning algorithm. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 358–363, 2019. doi: 10.1109/ICMLA.2019.00067. URL https://arxiv.org/abs/2002.03425. Author manuscript, arXiv v3, 2021.

Zebin Yang, Aijun Zhang, and Agus Sudjianto. GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192, 2021. doi: 10.1016/j.patcog.2021.108192. URL https://arxiv.org/abs/2003.07132.