NS-AIME: Rule-Guided Re-Constraint of Approximate Inverse Explanations

Takafumi Nakanishi[1,*]

[1]Tokyo University of Technology, Hachioji, Tokyo 192-0982, Japan

Abstract

Post-hoc explanation methods often face a tension between mathematical faithfulness and human-readable logical structure. This paper presents NS-AIME, a rule-guided re-constraint layer on top of Approximate Inverse Model Explanations (AIME). Starting from AIME's deterministic global attribution matrix, NS-AIME induces monotonicity, thresholds, and pairwise co-activation rules from data and solves a convex optimization problem that stays close to the original matrix while encouraging rule satisfaction. We report preliminary results on two vectorized datasets (Titanic and 20Newsgroups) across logistic regression, random forest, and LightGBM. Compared with raw AIME, NS-AIME substantially improves rule satisfaction, especially on sparse text data, while sometimes deviating strongly from the original attribution. To avoid masking problematic model behavior, we therefore advocate dual reporting: the raw AIME explanation and the constrained explanation are shown together, and large divergence is treated as an audit signal rather than as a silent replacement. The current evidence is limited to two datasets and no dedicated rule-learning baselines; this paper reports the method and initial findings and invites feedback on evaluation design.

Keywords

Explainable AI, approximate inverse explanations, neuro-symbolic explanations, rule-guided explanations, logical consistency

1. Introduction

Post-hoc explainability methods are widely used to interpret black-box predictors, but many of them emphasize local faithfulness and do not explicitly enforce a coherent global logic across explanations. Perturbation-based methods such as LIME and SHAP are flexible and influential, yet their outputs may vary with sampling, background choices, or locality assumptions [1, 2]. In contrast, Approximate Inverse Model Explanations (AIME) construct a deterministic inverse surrogate that maps model outputs back into feature space, yielding stable global attribution vectors without perturbation sampling [3]. The resulting explanation is mathematically well defined, but it does not necessarily follow monotonic or threshold-like relations that humans typically expect from rule-based reasoning.

This paper focuses on a narrow but practically important question: can a stable inverse explanation be re-constrained so that it becomes logically easier to audit, while still preserving a visible link to the original attribution? To this end, we propose NS-AIME, a post-hoc re-constraint layer applied to the AIME matrix. NS-AIME induces simple symbolic regularities directly from data—monotonic signs, feature thresholds, and pairwise co-activation rules—and then solves a convex optimization problem that encourages the explanation matrix to satisfy those rules while remaining close to the original AIME solution.

The contribution is intentionally scoped as late-breaking preliminary work. We do not claim that NS-AIME solves general explanation faithfulness for all model classes, nor do we claim superiority over dedicated rule-learning systems. Instead, we report initial evidence on two vectorized datasets and three classical model families, and we use the results to sharpen the methodological position of the

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.

nakanishi@stf.teu.ac.jp (T. Nakanishi)

https://takafuminakanishi.com/takafumi-nakanishi/ (T. Nakanishi)

0000-0003-1029-6063 (T. Nakanishi)

approach. In particular, this study makes four points explicit.

1. NS-AIME is a post-hoc explanation layer, not a new predictive rule learner. It operates on a trained black-box through its inverse surrogate.
2. The linearity lies in the inverse surrogate used by AIME; it is not an assumption that the original black-box model is globally linear.
3. The current implementation uses monotonicity, thresholds, and pairwise AND rules for tractability and readability; higher-order conjunctions and XOR-like interactions remain future work.
4. When the constrained explanation deviates strongly from raw AIME, this should be interpreted as an audit signal, not automatically as an improvement.

We further restrict our scope to trained classifiers on vectorized inputs $x \in \mathbb{R}^d$ such as tabular features or TF–IDF vectors. Dense continuous modalities such as raw images are outside the current empirical scope and would require an additional feature abstraction layer.

2. Related Work

Our work sits between post-hoc attribution methods and rule-based interpretable models. Local explanation methods such as LIME, SHAP, and Anchors estimate feature importance or rule conditions around a prediction and are standard model-agnostic baselines in XAI [1, 2, 4]. Broader surveys organize this design space and emphasize the distinction between local/global and intrinsic/post-hoc explanations [5].

On the rule-model side, scalable modern approaches include Scalable Bayesian Rule Lists (SBRL) [6], Interpretable Decision Sets (IDS) [7], CORELS [8], Bayesian Rule Sets [9], Falling Rule Lists [10], and post-processing methods such as QCBA [11]. NeuralLP provides a differentiable mechanism for learning logical rules, primarily in knowledge-base reasoning settings [12], while AMIE addresses Horn-rule mining in relational knowledge bases [13]. Historical model-extraction work such as TREPAN [14] and tree-ensemble interpretation methods such as inTrees [15] are also relevant because they recover concise symbolic structure from more complex predictors.

NS-AIME differs from these methods in one central respect: it does not replace the black-box with a rule-based predictor. Instead, it starts from a deterministic inverse explanation matrix obtained from a trained model and then imposes rule-guided structure on that matrix. In other words, rule learners optimize predictive rule sets, whereas NS-AIME optimizes the shape of an explanation produced by a separate model.

A separate naming issue also deserves clarification. Our base method is called AIME [3], whereas AMIE is a well-known framework for mining Horn rules from knowledge bases [13]. Despite the similar acronym, the problems are different: AMIE mines logical rules from symbolic relational data, while NS-AIME re-constrains feature-attribution matrices for post-hoc explanation of supervised classifiers.

3. Method

3.1. AIME as an inverse surrogate

Let $X \in \mathbb{R}^{n \times d}$ be the data matrix and let $\hat{Y} = f(X) \in \mathbb{R}^{n \times m}$ denote the predicted class probabilities of a trained classifier f. AIME defines a deterministic inverse surrogate as

$$A^{\dagger} = X^{\top}\hat{Y}(\hat{Y}^{\top}\hat{Y} + \varepsilon I_m)^{-1}, \tag{1}$$

where $\varepsilon > 0$ is a small ridge term for numerical stability. The k-th column of $A^{\dagger}$ is interpreted as the global contribution vector for class k.

Equation (1) should be read as a global linear inverse surrogate, not as an exact inverse of the original black-box. NS-AIME therefore improves the interpretability of this inverse surrogate, rather than claiming to reconstruct arbitrary nonlinear inverse dynamics.

3.2. Rule induction

From $A^\dagger$, the data matrix X, and the model outputs $\hat{Y}$, NS-AIME induces three classes of simple rules.

Monotonicity. For class k, each feature j receives a sign label

$$M_{jk} \in \{-1, 0, +1\},$$

where $+1$ means the contribution should be nonnegative, -1 means nonpositive, and 0 means unconstrained. The sign is induced only for sufficiently important features.

Thresholds. For each selected feature j and class k, a threshold t_{jk} is estimated from the empirical feature values of instances predicted as class k. For sparse text features, presence ($t_{jk} = 0$) is used.

Pairwise co-activations. For interpretability and tractability, we retain only pairwise conjunctions of high-contributing features:

$$(x_{j_1} \geq t_{j_1 k}) \wedge (x_{j_2} \geq t_{j_2 k}) \Rightarrow k,$$

with the direction flipped when the corresponding monotonicity is negative.

The restriction to monotonic, threshold, and pairwise AND rules is deliberate. It keeps the optimization convex and the explanations readable, but it does not capture higher-order or XOR-like interactions.

3.3. Rule-guided re-constraint

NS-AIME computes a new matrix $B \in \mathbb{R}^{d \times m}$ by solving

$$\min_B \ \|B - A^\dagger\|_F^2 + \alpha \mathcal{R}_{\text{mono}}(B; M) + \beta \mathcal{R}_{\text{rule}}(B; \mathcal{T}, \mathcal{P}), \tag{2}$$

where $\mathcal{R}_{\text{mono}}$ is a hinge penalty that discourages sign violations, and $\mathcal{R}_{\text{rule}}$ penalizes violated threshold and pairwise antecedents. The objective combines a quadratic proximity term with convex hinge penalties and can be optimized stably with gradient methods.

3.4. Dual reporting and audit safeguard

A central design choice is that the constrained explanation is not presented as a silent replacement of raw AIME. Instead, we advocate dual reporting:

1. show the original AIME matrix $A^\dagger$,
2. show the constrained matrix B, and
3. flag large divergence between them.

To make this explicit, we report two complementary quantities:

$$\text{RD}(B, A^\dagger) = \frac{\|B - A^\dagger\|_F}{\|A^\dagger\|_F}, \tag{3}$$

where lower values mean that B stays closer to raw AIME, and

$$\text{GA}(B, A^\dagger) = \frac{1}{m} \sum_{k=1}^{m} \cos\left(B_{\cdot,k}, A^\dagger_{\cdot,k}\right), \tag{4}$$

where higher values indicate stronger directional alignment. In earlier drafts, Eq. (3) was called "fidelity," which was misleading because it is a distance and lower is better. We therefore rename it relative deviation (RD).

This concern is related to the broader risk of rationalization or fairwashing, where a simpler surrogate can make a problematic black-box appear more acceptable than it is [16].

If RD is large or GA is low, the constrained explanation should be interpreted cautiously because it may be correcting the raw inverse surrogate too aggressively. This is particularly important in auditing settings, where one may want to reveal spurious or biased behavior rather than normalize it away.

4. Preliminary Experiments

4.1. Setup

We report preliminary results on two vectorized datasets: Titanic (tabular) and 20Newsgroups (sparse text). Three predictive model families were used: logistic regression, random forest, and LightGBM. The data were split into 70% training and 30% test partitions. The purpose of these experiments is narrow: they isolate the effect of the re-constraint layer on top of AIME. They are not intended as a comprehensive benchmark against rule-learning or rule-extraction baselines.

We use three metrics. Rule Satisfaction (RS) measures how often the induced rule antecedents are satisfied by the explanation matrix; higher is better. Relative Deviation (RD) is given by Eq. (3); lower is better. Global Alignment (GA) is given by Eq. (4); higher is better. RD is not bounded by 1: values above 1 simply mean that the constrained matrix departs from raw AIME by more than the Frobenius norm of the raw matrix itself.

Table 1
Preliminary comparison between raw AIME and NS-AIME. RS is higher-is-better, RD is lower-is-better, and GA is higher-is-better.

Dataset	Model	RS(raw)	RS(NS-AIME)	RD↓	GA↑
20Newsgroups	LightGBM	0.024	0.927	6.369	0.544
20Newsgroups	Logistic Reg.	0.053	0.916	5.273	0.473
20Newsgroups	Random Forest	0.152	0.750	6.429	0.352
Titanic	LightGBM	0.465	0.582	2.799	-0.216
Titanic	Logistic Reg.	0.467	0.597	2.007	-0.155
Titanic	Random Forest	0.469	0.582	3.096	-0.057

4.2. Observations

On 20Newsgroups, NS-AIME markedly increases RS across all three model families. The gains are largest for the models whose raw inverse explanations are least rule-consistent, which suggests that the re-constraint is effective in sparse high-dimensional settings where threshold and presence rules are natural.

On Titanic, the RS gains are smaller but consistent. At the same time, GA becomes negative for some models, which indicates a substantial directional change relative to raw AIME. Here we take a more cautious stance: a negative or very low GA should be regarded as an audit flag. It signals that the rule-guided explanation differs strongly from the inverse surrogate and therefore should be read alongside the raw explanation, not instead of it.

These preliminary experiments support the feasibility of the re-constraint idea, but they do not yet establish whether the resulting explanations are preferable to dedicated rule-based explanation systems or to human analysts in downstream tasks.

4.3. Qualitative examples on Titanic

Figure 1 shows the type of rule-structured explanation produced by NS-AIME on the Titanic dataset. Yellow nodes denote classes, gray nodes denote original features, and light-blue square nodes denote

induced pairwise AND relations. Edge colors indicate the sign or role of the relation: positive contributions are shown in green, negative contributions in red, co-occurrence links in gray, and edges involving AND nodes in blue.

For the `survived` class, the graph emphasizes feature patterns centered on `Sex_female`, class-related variables, and fare-related signals, together with pairwise conjunctions such as `AND(Sex_female^Pclass)` and `AND(Sex_female^Fare)`. For `not_survived`, the graph highlights a different arrangement centered on male-, fare-, and class-related structure. The purpose of this figure is qualitative: it illustrates how the re-constrained explanation becomes easier to inspect as a compact logical structure. At the same time, it should be interpreted together with the raw AIME explanation, since the graph visualizes the constrained explanation rather than the unconstrained inverse surrogate.

Dual-reporting view. Figure 2 makes the trade-off behind Table 1 concrete. On Titanic, NS-AIME does not merely shrink raw inverse attributions; it substantially redistributes weight across features. In both models, `pclass` becomes more dominant in the constrained view, while `fare` is weakened or changes sign relative to raw AIME, and the relative roles of `sex` and `age` also shift. This helps explain why RS improves whereas GA can become low or negative. Accordingly, large visual discrepancies such as those in Fig. 2 should be treated as an audit flag.

5. Discussion

The main strength of NS-AIME is conceptual: it isolates a clear middle layer between numerical attributions and symbolic rules. Starting from a stable inverse surrogate, it adds human-auditable structure without retraining the black-box. This is different from directly fitting an interpretable predictor and different from purely local perturbation methods, although it does not remove the broader concern that post-hoc explanations may be insufficient in high-stakes settings [17].

At the same time, the present evidence is narrow. First, the evaluation covers only two datasets and three classical model families; we do not report deep image or sequence models. Second, we do not yet compare against dedicated rule-learning or rule-extraction baselines such as SBRL, IDS, CORELS, or QCBA. Third, the method currently represents only monotonicity, thresholds, and pairwise AND relations; longer conjunctions, disjunction structure, and XOR-like interactions are outside the current optimization. Fourth, the current study does not include user-centered evaluation of whether the constrained explanations are easier to audit in practice.

A further limitation concerns data modality. The implementation is practical for vectorized and especially sparse inputs because the dominant term in Eq. (1) is the sparse product $X^\top \hat{Y}$. This does not imply that the method is already ready for dense pixel-level explanation. For dense continuous data, one would likely need feature grouping, embedding-level constraints, or other sparsification mechanisms before applying the present framework.

6. Conclusion

This paper presents NS-AIME as a rule-guided re-constraint layer for approximate inverse explanations. Preliminary results suggest that the method can substantially increase rule satisfaction, especially on sparse text data, while sometimes deviating strongly from the raw inverse surrogate. The key methodological point of this study is therefore not only the optimization itself, but also the proposed dual-reporting protocol: constrained explanations should be reported together with raw AIME, and large divergence should be treated as an audit signal.

Future work will focus on direct comparison with modern rule-learning baselines, evaluation on denser and more nonlinear settings, and extensions beyond pairwise monotonic rules.

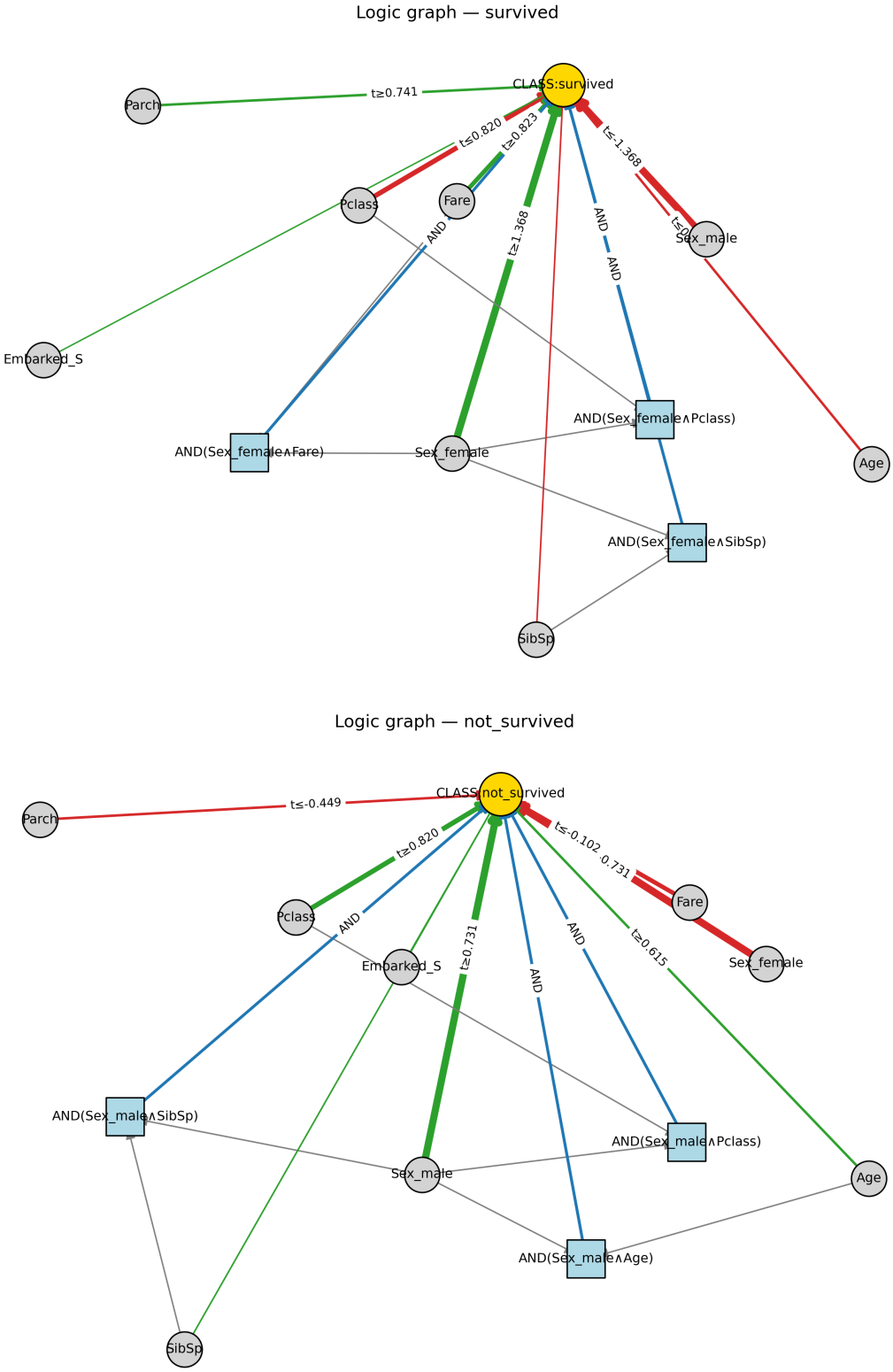
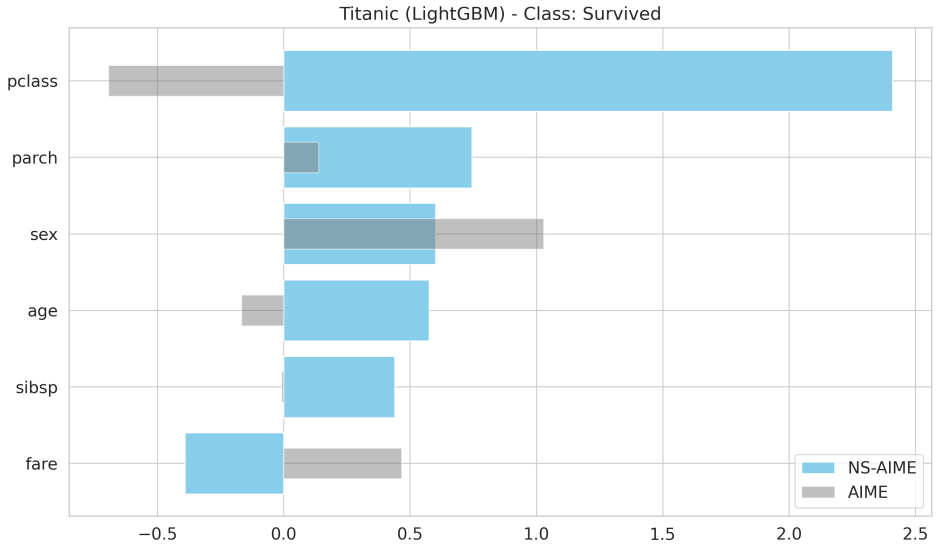


Figure 1: Qualitative logic graphs produced by NS-AIME for the Titanic dataset. Yellow nodes denote classes, gray nodes denote original features, and light-blue square nodes denote induced pairwise AND relations. Green edges indicate positive contributions, red edges indicate negative contributions, gray edges indicate co-occurrence links, and blue edges denote edges involving AND nodes. The figure illustrates the rule-structured view produced by the constrained explanation for `survived` and `not_survived`.

(a) LightGBM



(b) Logistic Regression

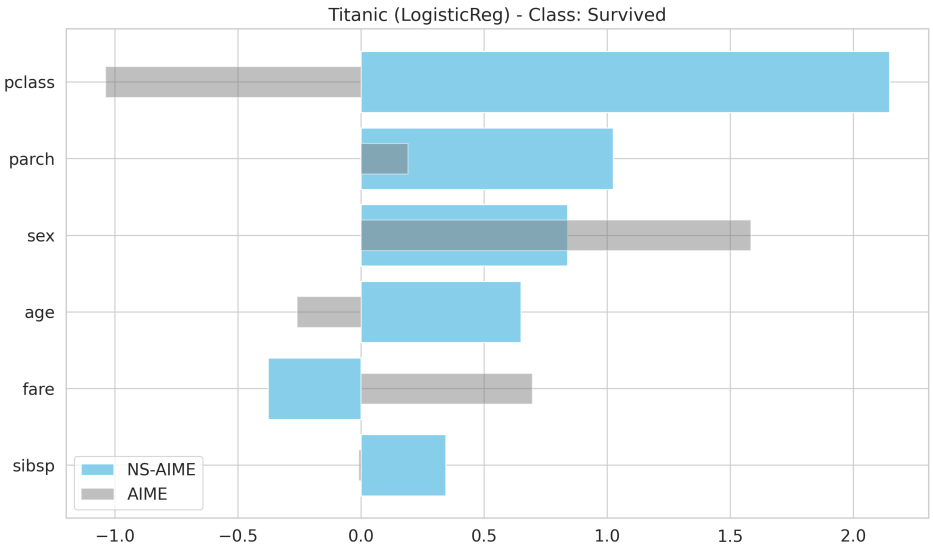


Figure 2: Dual-reporting on Titanic. Gray bars denote raw AIME and blue bars denote NS-AIME. Top: LightGBM. Bottom: Logistic Regression. Larger gaps correspond to higher RD / lower GA and should be read as audit signals.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT and DeepL in order to: grammar and spelling check, paraphrase and reword, translation support, and structural suggestions. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. doi:10.1145/2939672.2939778.

[2] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[3] T. Nakanishi, Approximate inverse model explanations (aime): Unveiling local and global insights in machine learning models, IEEE Access 11 (2023) 101020–101044. doi:10.1109/ACCESS.2023.3314336.

[4] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: high-precision model-agnostic explanations, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018, pp. 1527 – 1535.

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018). doi:10.1145/3236009.

[6] H. Yang, C. Rudin, M. Seltzer, Scalable Bayesian rule lists, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3921–3930.

[7] H. Lakkaraju, S. H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1675–1684. doi:10.1145/2939672.2939874.

[8] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning certifiably optimal rule lists for categorical data, Journal of Machine Learning Research 18 (2018) 1–78.

[9] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, P. MacNeille, A bayesian framework for learning rule sets for interpretable classification, Journal of Machine Learning Research 18 (2017) 1–37.

[10] F. Wang, C. Rudin, Falling Rule Lists, in: G. Lebanon, S. V. N. Vishwanathan (Eds.), Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, volume 38 of *Proceedings of Machine Learning Research*, PMLR, San Diego, California, USA, 2015, pp. 1013–1022.

[11] T. Kliegr, E. Izquierdo, Qcba: Improving rule classifiers learned from quantitative data by recovering information lost by discretisation, Applied Intelligence 53 (2023) 20797–20827. doi:10.1007/s10489-022-04370-x.

[12] F. Yang, Z. Yang, W. W. Cohen, Differentiable learning of logical rules for knowledge base reasoning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 2316–2325.

[13] L. A. Galárraga, C. Teflioudi, K. Hose, F. Suchanek, Amie: association rule mining under incomplete evidence in ontological knowledge bases, in: Proceedings of the 22nd International Conference on World Wide Web, WWW '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 413–422. doi:10.1145/2488388.2488425.

[14] M. W. Craven, J. W. Shavlik, Extracting tree-structured representations of trained networks, in: Proceedings of the 9th International Conference on Neural Information Processing Systems, NIPS'95, MIT Press, Cambridge, MA, USA, 1995, p. 24–30.

[15] H. Deng, Interpreting tree ensembles with intrees, International Journal of Data Science and Analytics 7 (2019) 277–287. doi:10.1007/s41060-018-0144-8.

[16] U. Aivodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, A. Tapp, Fairwashing: the risk of rationalization, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 161–170.

[17] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.