Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

BARBiE: An Associative Rule-Based Interactive Framework for Explaining Black-Box Model

Moriom Chowdhury Kumu[†,*], Iain Smith[†], Osmar Zaiane[†]
[†] University of Alberta

Abstract

Post-hoc explainable artificial intelligence is often provided as a product, typically in the form of static explanation such as a feature-importance ranking or a local surrogate explanation. In contrast, real-world decision workflows demand explanation as a process, characterized by interactivity in which users explore the decision output with what-if questions to develop understanding and trust. Existing explainers are often static, and their output is sensitive to how local samples around the instance are selected. Although rule-based local surrogates can expose feature interactions, user edits often require repeated resampling and retraining, limiting their usability for real-time what-if analysis. To address these gaps, we introduce **BARBiE**, a model-agnostic framework for instance-level explanation that integrates an association-rule surrogate with an interactive interface. For a given query instance, BARBiE constructs an instance-centered neighborhood, queries the black-box model for labels, and trains a compact association-rule surrogate. Explanations are provided only when the surrogate output matches the black-box decision for the query instance. BARBiE presents IF–THEN rules with support, confidence, and a p-value from Fisher's exact test. In addition, BARBiE computes rule-grounded, signed feature importance by aggregating instance-aware contributions from the rule base. Importantly, BARBiE supports quick what-if analysis without resamples and retraining the surrogate model. Across four tabular datasets and a user study, we evaluated BARBiE against LIME, SHAP, and BARBE using user ratings of informativeness, understandability, trustworthiness, and satisfaction. Across tasks, BARBiE consistently received higher ratings than the baselines, providing supports that process-centric interactive explanations improve informativeness and understandability and contribute to higher trust and user satisfaction.

Keywords: XAI, BARBiE, associative classifier, rule-based method, post-hoc method.

1. Introduction

Machine learning (ML) models are increasingly being deployed to support decision-making in critical domains, including finance, healthcare, and public services [1]. In such settings, predictive accuracy is insufficient; stakeholders want to understand why a model produced a particular outcome to assess reliability, ensure accountability, and build trust. Nevertheless, many high-performing models operate as black boxes, restricting transparency and human understanding [2–4]. Post-hoc eXplainable Artificial Intelligence (XAI) methods aim to explain black-box model behavior without requiring access to or modification of the model's internal structure [5]. Existing approaches, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) attribute predictions to input features through local surrogate modeling or Shapley-value approximations [6–8]. Although these methods effectively highlight influential features, they are often presented as static *explanation products*: a fixed explanation for a single input, with outputs that can vary depending on how local samples are generated and which reference sample data are chosen [9].

Recent research work identifies a deeper conceptual mismatch: the explanation is often presented as a *product* rather than an ongoing *process* [10–13]. As a result, many tools

* kumu@ualberta.ca

produce explanations that seem interpretable at first glance but do not support user interaction. In practice, decision-makers rarely stop at a single question such as "Why this prediction?" Instead, they engage in an iterative inquiry process, probing the model with *what-if* questions such as "What if this feature changed?" or "What conditions are driving this outcome?" Each response helps them improve their understanding and trust. Methods that do not support this interactive back-and-forth may therefore struggle to foster trust and understanding, even when the explanation itself is faithful [14, 15].

Rule-based explanations align with human decision logic by expressing local behavior through human-readable IF–THEN statements which can naturally capture feature interactions [16–19]. Local rule-based explainers, such as LOcal Rule-based Explanations (LORE) [6], and association-rule surrogates, such as Black-box Association Rule-Based Explanations (BARBE) [20], induce rule sets from locally sampled data labeled with black-box. These methods improve interpretability by proving structured rules, but treat the explanation as a *fixed output* rather than an interactive workflow. Explanations are generated for one fixed input, and user edits typically require regenerating the samples and retraining the surrogate. This slows interactive exploration and scenario testing, which are crucial for improving user trust and satisfaction. Interactive XAI incorporates user-driven *what-if* edits, enabling users to explore model behavior [21, 22]. Previous systems often highlight visualizations, feature scores, or counterfactuals, but few provide interactive rule-based explanations.

This paper introduces **Black-box Association Rule-Based Interactive Explanations (BARBiE)**, a model-agnostic framework that generates compact IF–THEN rules and supports user-driven *what-if* analysis through an interactive interface. BARBiE constructs an instance-centered local neighborhood, labels it using the pretrained black-box model, and trains an associative classifier using SigD2 [23] to generate a compact set of IF–THEN rules with support, confidence, and a p-value from Fisher's exact test. In addition, it enforces an instance-level *fidelity gate* that returns explanations only when the surrogate matches the black-box decision. Notably, BARBiE enables fast *what-if* analysis without retraining by reusing the learned rule set and updating explanations through rule applicability checks.

BARBiE is evaluated on four tabular datasets (Loan Approval, Iris, Glass, and Wine) and compared against LIME, SHAP, and BARBE in a controlled user study (n=34) guided by the DARPA XAI evaluation principles [24]. Participants rated each method on a 5-point Likert scale in four dimensions: *informativeness*, *understandability*, *trustworthiness*, and *satisfaction*. BARBiE received higher ratings than all baselines in these dimensions, supporting the claim that interactive explanations improve perceived informativeness and understandability while increasing trust and satisfaction. The contributions of this paper are as follows:

- We extend BARBE into **BARBiE**, a model-agnostic interactive explanation framework for tabular classifiers and *what-if* analysis.
- We introduce an instance-level *fidelity gate* that returns explanations only when the surrogate matches the black-box decision on the query instance.
- We show that BARBiE improves perceived informativeness, understandability, trustworthiness, and satisfaction relative to LIME, SHAP, and BARBE.

2. Methodology

Given a trained black-box model $f : \mathbb{R}^d \to [0, 1]$ and a query instance $\mathbf{x} \in \mathbb{R}^d$, BARBiE builds a compact association-rule surrogate around $\mathbf{x}$. In addition, BARBiE enforces an instance-level fidelity check, and supports rapid *what-if* exploration without new training.

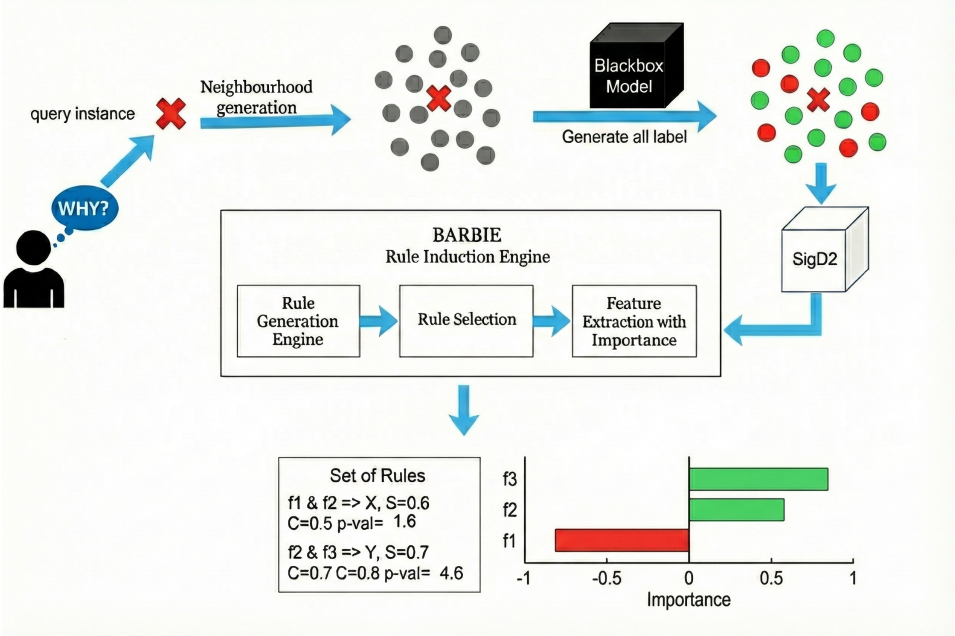

Figure 1. BARBiE pipeline: For a query instance, neighborhood generation and labels, train surrogate model, then generate, selects and returns ranked IF–THEN rules, with feature importance to produces explanations for an instance.

Figure 1 illustrates the BARBiE pipeline: input query, neighborhood generation and labeling, train surrogate model, generate rule, select, and return ranked rules with feature importance.

2.1. **Data Preparation and Discretization**

The efficacy of BARBiE, particularly in generating rule-based explanations, is highly dependent on how data features are handled. BARBiE begins with dataset preparation, executed once per dataset to ensure consistent encodings. Features are partitioned into categorical and numeric types. Numeric features are discretized into a small number of intervals, and categorical features are encoded using one-hot encoding.

2.2. **Neighborhood Generation**

To explain the prediction $f(\mathbf{x})$, BARBiE constructs a synthetic neighborhood $\mathcal{N}$ around the query instance $\mathbf{x}$. Perturbations are sampled in the encoded feature space using distribution-aware generators. By default, samples are drawn from a multivariate normal distribution, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ is the feature covariance. Each perturbed point $\mathbf{z}_i$ is labeled using the black-box model, yielding a locally labeled dataset $\{(\mathbf{z}_i, f(\mathbf{z}_i))\}_{i=1}^{m}$ used to train the interpretable surrogate. Unlike prior work that uses independent Gaussian perturbations, BARBiE samples from a multivariate normal distribution so correlated features are perturbed jointly, producing more realistic neighborhoods and higher-quality surrogates [25].

2.3. **Interpretable Surrogate Learning**

BARBiE trains an interpretable surrogate model $g : \mathbb{R}^d \to [0, 1]$ to the locally labeled neighborhood using SigD2 [23], a statistically significant associative classifier that incorporates two-stage pruning. Compare with SigDirect [26], SigD2 uses two-stage pruning to remove redundant, low-utility rules, producing a more compact and readable local explanation set. SigD2 surrogate generates IF–THEN rules $r : A \to y$, where A is a set of discretized feature conditions and y is the predicted class label. Rules are annotated with support $\text{supp}(r)$, confidence $\text{conf}(r)$, and a bounded significance score derived from Fisher's exact test. Conservative thresholds on rule length, support, and confidence keep the rule set compact [16]. Two-stage pruning eliminates redundant and low-utility rules, yielding a compact, interpretable local rule base. In contrast to generic trees or linear surrogates, SigD2 provide a compact and readable rule set that approximates the behavior of f in the vicinity of $\mathbf{x}$.

2.4. **Local Fidelity Gate**

To ensure a minimum threshold of faithfulness, BARBiE introduces a *local fidelity gate* to quantify how well the surrogate $g(\mathbf{x})$ match $f(\mathbf{x})$ within a local context. A mismatch ($g(\mathbf{x}) \neq f(\mathbf{x})$) indicates that the surrogate has not accurately captured the local decision boundary relevant to $\mathbf{x}$. If the check fails, BARBiE resamples the neighborhood and retrains the surrogate up to a bounded number of attempts. Only when the fidelity gate satisfies the check ($g(\mathbf{x}) = f(\mathbf{x})$), BARBiE finalizes the current surrogate: rules are ranked and decoded, applicable rules are selected, and feature attributions are computed and returned to the user. This mechanism prevents unfaithful explanations from being displayed and ensures that the extracted rules and attributions reflect the black-box decision for the specific instance under analysis.

2.5. **Rule Extraction and Feature Importance**

Once fidelity is confirmed ($g(\mathbf{x}) = f(\mathbf{x})$), BARBiE extracts the local rule base $\mathcal{R}$, returns the all rule set for context, and selects the subset $\mathcal{R}(\mathbf{x}) \subseteq \mathcal{R}$ satisfied by $\mathbf{x}$. Rules are decoded into human-readable predicates and ranked by agreement with $f(\mathbf{x})$, then by confidence, support, and p-value. BARBiE derives per-instance feature importance directly from the surrogate's rule base. For each feature, BARBiE aggregates signed evidence from all rules that include the feature. Each feature contribution is weighted by rule quality (support, confidence, p-value) and by the degree to which the rule is satisfied by $\mathbf{x}$. rule weight $s(r) = \text{supp}(r) \cdot \text{conf}(r) \cdot \phi(r)$, BARBiE computes per-instance signed importance for feature j by aggregating contributions across rules that contains feature j and then applies ℓ_1 normalization. A feature is important when it included in high-quality rules (high support, confidence, and p-value) that are largely satisfied by x and whose labels align with the predicted class. Signed scores are positive when the consequent label of a rule aligns with the predicted class, and negative otherwise.

3. **Results and Discussion**

Our evaluation centers on explanation as an interactive *what-if* process, addressing a key gap in interpretability: whether an explainer supports interactive inquiry to improve user understanding, trust and satisfaction rather than a single static output. We evaluate BARBiE on four widely used tabular datasets containing numeric and categorical attributes: the Loan Approval dataset [27], and the Iris, Glass, and Wine datasets from the UCI Machine Learning Repository [28]. We compare BARBiE against LIME, SHAP, and BARBE

through a controlled user study measuring user-rated informativeness, understandability, trustworthiness, and satisfaction.

3.1. **User Study Results**

To assess human-centered outcomes, we conducted a controlled user study aligned with DARPA XAI evaluation guidance [24], evaluating explanations using user-rated informativeness, understandability, trustworthiness, and satisfaction. This user study tests not only whether participants found the explanation *outputs* useful, but also whether interactive (what-if edits) improves informativeness, understandability, trustworthiness, and satisfaction. A total of 34 participants completed the study, with ages ranging from 18 to 50 (mean = 26.02, SD = 6.42); 46% identified as female. Participants reported mixed familiarity with machine learning (60% "very familiar", 30% "somewhat familiar" or "neutral").

Participants compared BARBiE against LIME, SHAP, and BARBE using anonymous, within-task evaluations. Participants evaluated explanations using four research questions (5-point Likert scale) across four dimensions: *Informativeness*, *Understandability*, *Trustworthiness*, and *Satisfaction*. Four research questions as follows :

- **RQ1 (Informativeness):** Are the rankings or rules provided by association rule-based methods effective in conveying useful information to users?
- **RQ2 (Understandability):** Does allowing users to interact with systems and modify input features enhance their understanding and trust in the AI system?
- **RQ3 (Trustworthiness):** Do association rule-based explanations help users trust the AI system more than existing post-hoc methods?
- **RQ4 (Satisfaction):** Does interactivity (rule visualization + what-if analysis) improve user satisfaction relative to static feature-importance explanations?

For each task, participants inspected the explanation for a prediction and then provided Likert ratings. Participants evaluated the model's explanations for each prediction before providing their responses on a Likert scale. Figure 2 shows the distribution of participant responses across the four evaluation dimensions—Informativeness, Understandability, Trustworthiness and Satisfaction. Table 1 provides a statistical summary of the user feedback, reporting the mean and standard deviation (M±SD) for each research question.

BARBiE received strong positive ratings on **RQ1 (Informativeness)**, with a mean score of 4.07 ± 0.68. The distribution in Figure 2 shows that most responses were positive: 56% "Agree" and 26% "Strongly Agree" (82% total). This suggests that participants found BARBiE's displayed rule conditions and rule-grounded attributions to be informative and sufficient for understanding the basis of the model's decision.

For **RQ2 (Understandability)**, BARBiE achieved 3.74 ± 0.66, with 63% "Agree" and 7% "Strongly Agree" (70% total positive). While 26% selected "Neutral" and 4% "Disagree", the overall pattern indicates that most users could follow the explanation logic and, importantly, benefited from being able to probe the decision through what-if edits—consistent with Watson's [10] claim that interpretability must support interactive inquiry rather than only static outputs.

On **RQ3 (Trustworthiness)**, BARBiE also achieved 3.74 ± 0.76. A majority of participants reported trust-positive responses (48% "Agree", 15% "Strongly Agree"; 63% total positive). The remaining 33% "Neutral" and 1% "Disagree" highlight a common reality in applied XAI: trust calibration is gradual and depends on repeated exposure and domain context. Nevertheless, the overall trend supports the argument that an interactive rule substrate helps users assess whether model behavior is stable under plausible changes.

Finally, **RQ4 (Satisfaction)** was notably high (4.00 ± 0.78), with 41% "Agree" and 30% "Strongly Agree" (71% total positive). This suggests that participants valued the overall

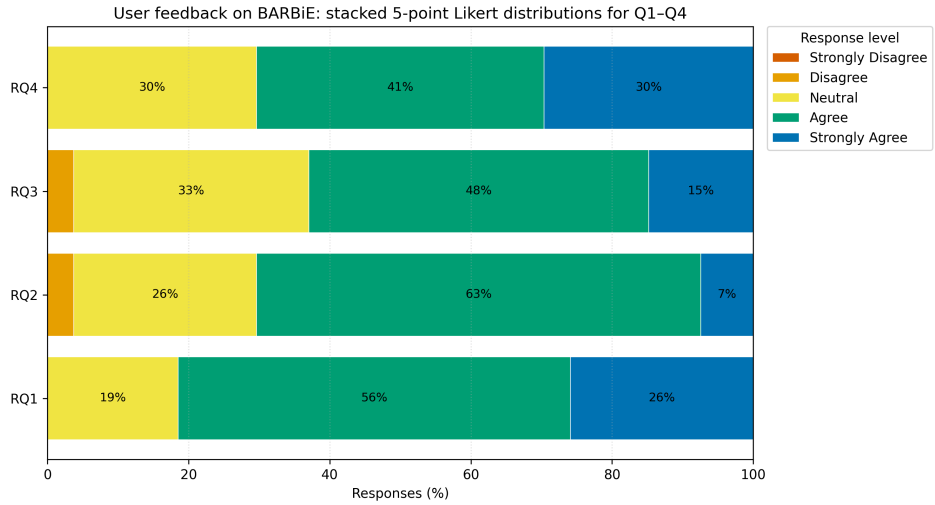

Figure 2. User feedback on BARBiE: stacked 5-point Likert distributions for RQ1–RQ4 (Informativeness, Understandability, Trustworthiness, Satisfaction). Each bar shows the percentage selecting Strongly Disagree → Strongly Agree.

Table 1. User study results (mean±sd, 5-point Likert scale).

Method	Info	Understand	Trust	Satisfaction
LIME	3.6±0.5	3.3±0.2	3.2±0.5	3.4±0.3
SHAP	3.7±0.4	3.4±0.3	3.3±0.4	3.5±0.4
BARBE	3.8±0.5	3.5±0.4	3.5±0.1	3.8±0.3
BARBiE	**4.07**±0.68	**3.74**±0.66	**3.74**±0.76	**4.00**±0.78

experience of using BARBiE as an exploratory tool—i.e., not just reading an explanation, but *working with it* to test scenarios and understand decision conditions.

As shown in Table 1, BARBiE achieves the highest mean ratings across all four dimensions. Participants rated BARBiE as more informative, easier to understand, more trustworthy, and more satisfying than existing feature-importance-based explainers and static rule-based baselines. These results suggest that showing compact IF–THEN rules with interactive *what-if* capabilities improves users' understanding of model logic and trust.

3.2. **Statistical Significance**

To compare BARBiE with each baseline, we use paired one-sided t-tests because the study follows a within-subject design (each participant rated BARBiE and the same baseline methods). For each participant i, we compute paired differences $d_i = s_i(\text{BARBiE}) - s_i(\text{baseline})$ and test $H_0 : \mathbb{E}[d] \leq 0$ against $H_1 : \mathbb{E}[d] > 0$. To control the family-wise error rate across the multiple baseline comparisons within each dimension, we apply the Holm–Bonferroni correction [29] with $\alpha = 0.05$.

As shown in Table 2, **BARBiE** significantly outperformed the industry-standard baselines (LIME and SHAP) on all four research questions. The largest gains were observed for *Satisfaction*: **BARBiE** achieved a mean score of $M = 4.00$ and showed a highly significant improvement over both LIME ($t = 4.65, p_{adj} < 0.05$) and SHAP ($t = 4.23, p_{adj} < 0.05$). For *Informativeness*, **BARBiE** was the only method with a mean above 4.0 and it also significantly outperformed BARBE ($p_{adj} = 0.048$). **BARBiE** further achieved higher mean scores

than BARBE for *Understandability* ($M = 3.74$ vs. 3.50) and *Trustworthiness* ($M = 3.74$ vs. 3.50), but these differences are not significant. Overall, the results indicate a consistent trend toward improved user experience with **BARBiE**.

Table 2. Paired t-test results comparing BARBiE to baselines across four research questions. P-values are one-tailed and adjusted using the Holm-Bonferroni method.

Question	Baseline	Mean Diff	t-stat	p_{adj}	Sig.
RQ1: Info	LIME	0.63	4.76	< 0.01	**
	SHAP	0.22	2.28	< 0.05	*
	BARBE	0.26	1.76	< 0.05	*
RQ2: Understand	LIME	0.52	4.19	< 0.01	**
	SHAP	0.41	3.70	< 0.05	*
	BARBE	0.26	1.66	$>$ **0.05**	ns
RQ3: Trust	LIME	0.59	4.84	< 0.01	**
	SHAP	0.44	4.00	< 0.01	**
	BARBE	0.33	1.80	$>$ **0.05**	ns
RQ4: Satisfaction	LIME	0.52	4.65	< 0.01	**
	SHAP	0.41	4.23	< 0.01	**
	BARBE	0.30	1.69	$>$ **0.05**	ns

Note: * $p < .05$, ** $p < .01$, *ns: not significant.*

4. Discussion

BARBiE distinguishes itself from prior rule-based explainers by bridging the gap between static explanation and the iterative "process" of human reasoning. This is achieved through two primary innovations: interactivity and an improved surrogate model.

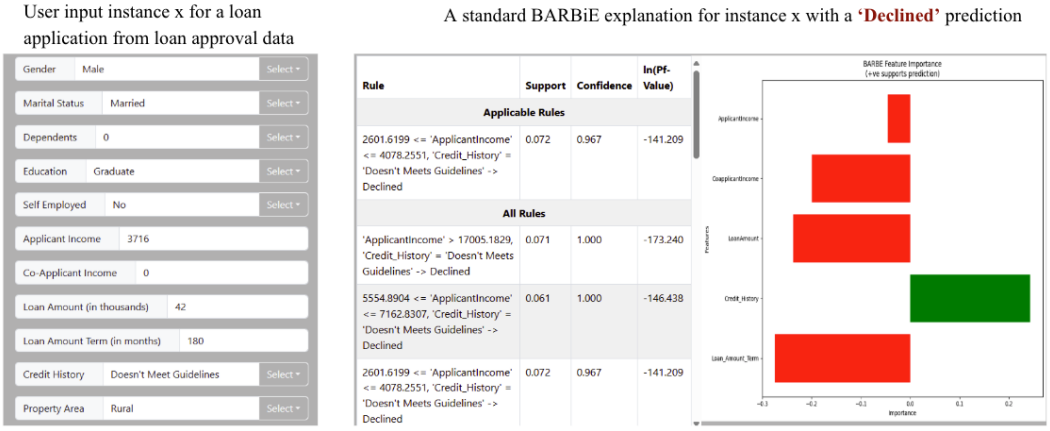

Figure 3. A BARBiE explanation for an instance with a *Declined* prediction, showing the dominant rule conditions.

Firstly, BARBiE supports interactive *what-if* analysis by updating rule applicability and feature importance without retraining the black-box model or the surrogate. This enables users to iteratively explore how changes to input features affect both predictions and explanations. Figure 3 illustrates BARBiE's core capability: supporting explanation as an *interactive process*. For an applicant instance initially predicted *Declined*, BARBiE presents

a set of applicable IF–THEN rules. These rules, backed by support, confidence and p-value, reveal that the primary reason for the "Decline" decision is a `Credit_History` that fails to meet guidelines. To support "what-if" process, BARBiE allow feature modifications—such as updating `Credit_History` to meet guidelines. BARBiE immediately updates the explanation: the old rule is removed and new rules is presented as applicable rules supporting an *Approved* prediction (Figure 4). BARBiE select the new applicable rules from all rules without retraining the surrogate model. This interaction is precisely the kind of inquiry loop emphasized by Watson [10]: users do not trust a static explanation; they test the decision output by proposing changes and observing whether the resulting explanation remains consistent and reasonable under new conditions.

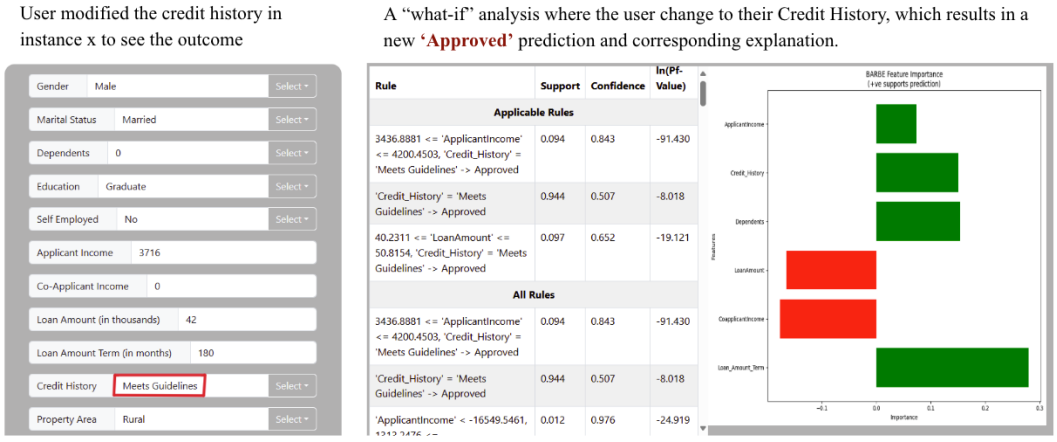

Figure 4. A *what-if* analysis where modifying `Credit_History` results in an *Approved* prediction and an updated rule-based explanation.

Secondly, Interactive explanations can be persuasive even when they are not faithful. BARBiE mitigates this risk with an instance-level fidelity gate that returns explanations only when the surrogate matches the black-box decision in the query instance. This provides a safety constraint on the process: users do not iterate on an explanation substrate that already disagrees with the model at the starting point. Furthermore, while BARBE introduced SigDirect, an association rules based surrogate model, it often suffered from redundancy and high cognitive load. In contrast, BARBiE's transition to the **SigD2** surrogate with a two-stage pruning strategy prioritizes concise, non-redundant IF–THEN rules [23]. Furthermore, unlike rule based method LORE, BARBiE provides rigorous statistical metrics—including support, confidence, and p-values—ensuring that the insights provided are not only readable but statistically significant.

5. Conclusion

This paper presented **BARBiE**, a model-agnostic framework for explaining black-box predictions on tabular data via *interactive*, *rule-based* explanations. Motivated by *product vs. process* critique, BARBiE treats interpretability as an interactive workflow rather than a fixed output. For a given query instance, it generates a compact local rule set with SigD2 and two-stage pruning, filters unreliable explanations using an instance-level fidelity gate, and support *what-if* analysis. Empirically, BARBiE produced concise IF–THEN rules that make conditional logic and feature interactions explicit. In a controlled user study, BARBiE achieved the highest mean ratings on informativeness, understandability, trustworthiness,

and satisfaction relative to LIME, SHAP, and BARBE. Paired one-sided tests further indicate that BARBiE provides a measurable informativeness advantage over attribution-based explanation products, consistent with the claim that process support is a key missing component in deployed XAI.

Acknowledgements

Part of this work has taken place in the Alberta Machine Intelligence Institute (Amii) xAI Lab at the University of Alberta. The authors sincerely thank the participants in the user study for generously volunteering their time and effort. We also used a state-of-the-art large language model to improve the clarity and writing quality of text in this paper.

References

[1] C. Acun and O. Nasraoui. "Pre Hoc and Co Hoc Explainability: Frameworks for Integrating Interpretability into Machine Learning Training for Enhanced Transparency and Performance". In: *Applied Sciences* 15.13 (2025), p. 7544.

[2] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain. "Interpreting black-box models: a review on explainable artificial intelligence". In: *Cognitive Computation* 16.1 (2024), pp. 45–74.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[4] Z. C. Lipton. "The Mythos of Model Interpretability". In: *Communications of the ACM* 61.10 (2018), pp. 36–43. DOI: 10.1145/3233231.

[5] B. Finzel. "Current methods in explainable artificial intelligence and future prospects for integrative physiology". In: *Pflügers Archiv-European Journal of Physiology* 477.4 (2025), pp. 513–529.

[6] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. "Local rule-based explanations of black box decision systems". In: *arXiv preprint arXiv:1805.10820* (2018).

[7] L. S. Shapley et al. "A value for n-person games". In: (1953).

[8] F. Doshi-Velez and B. Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[9] M. Motallebi and et al. "Explaining Black-Box Models with Association Rules". In: *Machine Learning* (2023).

[10] D. S. Watson. "Conceptual challenges for interpretable machine learning". In: *Synthese* 200.2 (2022), p. 65.

[11] T. Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[12] A. Bertrand, T. Viard, R. Belloum, J. R. Eagan, and W. Maxwell. "On Selective, Mutable and Dialogic XAI: A Review of What Users Say about Different Types of Interactive Explanations". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. 2023. DOI: 10.1145/3544548.3581314.

[13] Q. V. Liao, D. Gruen, and S. Miller. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. 2020, pp. 1–15. DOI: 10.1145/3313831.3376590.

[14] E. Gagnon, A. de Regt, and L. LaPointe. "Clarity in Complexity: Advancing AI Explainability through Sensemaking". In: *Proceedings of the 58th Hawaii International Conference on System Sciences, HICSS 2025*. 2025, pp. 1400–1409.

[15] D. Hemment, D. Murray-Rust, V. Belle, R. Aylett, M. Vidmar, and F. Broz. "Experiential AI: Enhancing explainability in artificial intelligence through artistic practice". In: (2022).

[16] B. Liu, W. Hsu, and Y. Ma. "Integrating Classification and Association Rule Mining". In: *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD)*. 1998.

[17] X. Yin and J. Han. "CPAR: Classification based on Predictive Association Rules". In: *Proceedings of the SIAM International Conference on Data Mining (SDM)*. 2003.

[18] J. H. Friedman and B. E. Popescu. "Predictive Learning via Rule Ensembles". In: *The Annals of Applied Statistics* 2.3 (2008). DOI: 10.1214/07-AOAS148.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018.

[20] M. Motallebi, M. T. A. Anik, and O. R. Zaïane. "Explaining decisions of black-box models using barbe". In: *International Conference on Database and Expert Systems Applications*. Springer. 2023, pp. 82–97.

[21] M. Chromik. "reshape: A framework for interactive explanations in xai based on shap". In: (2020).

[22] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. "explAIner: A visual analytics framework for interactive and explainable machine learning". In: *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 1064–1074.

[23] N. Sood and O. Zaiane. "Building a competitive associative classifier". In: *International Conference on Big Data Analytics and Knowledge Discovery*. Springer. 2020, pp. 223–234.

[24] D. Gunning and D. Aha. "DARPA's explainable artificial intelligence (XAI) program". In: *AI magazine* 40.2 (2019), pp. 44–58.

[25] I. N. Smith and O. R. Zaïane. "Faithful Perturbations and Evaluations for Post-Hoc Local Explanation Methods". In: *The 38th Canadian Conference on Artificial Intelligence* (2025).

[26] J. Li and O. R. Zaiane. "Exploiting statistically significant dependent rules for associative classification". In: *Intelligent data analysis* 21.5 (2017), pp. 1155–1172.

[27] D. Chatterjee. *Loan prediction problem dataset*. Kaggle. Mar. 2019. URL: https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset.

[28] M. Kelly, R. Longjohn, and K. Nottingham. *The UCI Machine Learning Repository*. https://archive.ics.uci.edu. University of California, Irvine, School of Information and Computer Sciences. 2024.

[29] S. Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70. DOI: 10.2307/4615733.