Explaining Black-Box AI Models with BARBiE: An Associative Rule-Based Interactive Framework

Abstract

Artificial Intelligence (AI) models are regularly deployed to support decision-making in critical domains. However, the opacity of black-box models limits user understanding, accountability, and calibrated trust. Existing post-hoc explainers primarily provide static feature-importance scores that are sensitive to neighborhood design and offer limited support for exploratory reasoning. Moreover, they lack interactivity, restricting users' ability to examine alternative scenarios.

To address these gaps, we introduce Black-box Association Rule-Based Interactive Explanations (BARBiE), a model-agnostic framework that couples an association-rule surrogate with an interactive interface for instance-level explanation. Without requiring access to model internals, BARBiE constructs an instance-centered local neighborhood and trains an associative classifier with two-stage pruning to produce compact IF–THEN rules annotated with support, confidence, and significance metrics. To ensure faithfulness at the instance level, BARBiE uses a fidelity gate that returns explanations only when the surrogate agrees with the black-box decision on the query instance. BARBiE enables interactive what-if analysis: as users modify input features, explanations update quickly by recomputing applicable rules and rule-grounded feature importance, without retraining the surrogate model. Feature importance is derived directly from the rule base by aggregating signed, instance-aware contributions weighted by rule quality and degree of rule satisfaction.

We evaluate BARBiE on four tabular datasets (Loan Approval, Iris, Glass, and Wine) and compare it against LIME, SHAP, and Black-box Association Rule-Based Explanations (BARBE). In a comprehensive user study (n=34), BARBiE receives higher ratings for informativeness, understanding, trust, and satisfaction than all baselines. Despite certain drawbacks related to counterfactual generation and rule growth, BARBiE's interactive, rule-based, and API-only design makes it a useful solution for explaining black-box models in enterprise settings.

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

1. Introduction

Machine learning (ML) models are increasingly deployed to support decision-making in critical domains such as finance, healthcare, and public services [1]. In these settings, model's predictive accuracy alone is not sufficient. Stakeholders wants to understand why a model produced a particular outcome in order to assess reliability, and ensure accountability. However, many deep ML-models operate as black boxes, offering limited transparency into their decision logic and restricting meaningful human oversight [2].

Post-hoc eXplainable Artificial Intelligence (XAI) methods aim to address this challenge by explaining a trained black-box model using only its input–output behavior, without access to or modification of model internals. Widely adopted approaches such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) assign importance scores to input features for a given prediction [3–5]. While effective at highlighting influential features, these methods are inherently static: explanations are generated for a fixed input and do not support user-driven exploration. As a result, they offer limited support for *what-if reasoning*, where users seek to understand how changes to input

*corresponding_author@example.ca

This article is © 2026 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author(s) identified above.

features would alter a model's decision. Moreover, feature-importance methods primarily emphasize additive effects and often fail to expose higher-order interactions or conditional logic that are central to human reasoning in many real-world decision processes.

Rule-based explanation methods provide a complementary perspective by expressing model behavior through human-readable IF–THEN statements. Approaches such as LOcal Rule-based Explanations (LORE) [3] and Black-box Association Rule-Based Explanations (BARBE) [6] demonstrate that compact rule surrogates trained on locally sampled, black-box-labeled data can capture feature interactions and present decision logic in an interpretable form. However, existing rule-based explainers remain largely non-interactive. Explanations are generated for a single, fixed input, and any modification to feature values typically requires regenerating neighborhoods or retraining the surrogate model. This computational overhead limits their practicality for exploratory analysis and scenario testing, which are critical for building calibrated trust and understanding.

This paper addresses this gap by introducing a framework that delivers *interactive*, *rule-based*, and *model-agnostic* explanations under realistic deployment constraints. We propose **Black-box Association Rule-Based Interactive Explanations (BARBiE)**, a framework designed for tabular classification settings where model internals are not available. BARBiE constructs an instance-centered local neighborhood, labels it using the black-box predictor, and trains a statistically disciplined association-rule surrogate using a two-stage pruning strategy. The resulting explanations consist of a small set of human-readable IF–THEN rules annotated with support, confidence, and capped statistical significance. To guard against misleading explanations, BARBiE employs a lightweight local fidelity gate that returns explanations only when the surrogate agrees with the black-box decision at the query instance.

Beyond static explanation, BARBiE enables real-time *what-if* analysis. Users can modify input features and immediately observe updated rules and rule-grounded feature importance, without retraining the surrogate model or re-mining rules. Feature importance is derived directly from the rule base by aggregating signed, instance-aware contributions, ensuring that numerical attributions remain explicitly tied to the displayed decision logic.

We evaluate BARBiE on four tabular datasets (Loan Approval, Iris, Glass, and Wine) and compare it against LIME, SHAP, and BARBE. Through a controlled user study, BARBiE receives higher ratings for informativeness, understanding, trust, and satisfaction than all baselines. These results demonstrate that interactive, rule-based explanations offer a practical and human-centered approach to understanding black-box predictions. This paper makes the following contributions:

- We propose **BARBiE**, a model-agnostic framework that combines association-rule learning with an interactive interface for explaining black-box predictions in supervised tabular classification.
- We introduce an instance-level *fidelity gate* that ensures explanations are returned only when the rule-based surrogate matches the black-box decision at the query instance.
- We enable efficient, real-time *what-if* analysis by updating rule applicability and rule-grounded feature importance without retraining the surrogate model or re-mining rules.
- We empirically show through a user study that BARBiE improves perceived informativeness, understanding, trust, and satisfaction compared to LIME, SHAP, and BARBE.

2. Related Work

XAI comprises methods that make complex machine-learning model behavior understandable—and actionable—for human stakeholders by producing faithful and useful explanations of the model's logic and decisions [7]. Implementation and deployment in critical fields like healthcare, finance, and the public sector often require transparency, understanding, accountability, and trust. Policy and governance efforts—notably the DARPA XAI program—formalize this need by promoting techniques that are faithful to a model's logic and useful for decision-making [8].

Explainable artificial intelligence (XAI) research has developed diverse techniques to improve the transparency of black-box models [9]. Prior work [2, 9, 10] mostly falls into two areas: model-specific methods and model-agnostic methods. Model-specific methods exploit a model's internal structure to attribute predictions. For example, Saliency Maps, Integrated Gradients, and Grad-CAM highlight influential inputs in text, tabular, or image settings [11–13]. While these methods leverage architecture-specific structure, they do not generalize to arbitrary black boxes and require sanity checks to ensure that attributions reflect learned representations rather than dataset priors, avoiding the conflation of visual plausibility with causal faithfulness [14, 15].

In contrast, Post-hoc, model-agnostic explainers like LIME and SHAP are widely used in XAI [4, 16–19]. They explain individual predictions by assigning importance scores to input features, using surrogate models or value approximations. These methods emphasize salient features but are usually static, providing a single snapshot per instance [20]. Existing work [21–24] shows explanations can be unstable due to sampling, kernel, or background choices, limiting their support for actionable reasoning. Additive features can mask feature interactions and do not naturally enable user-driven what-if analysis, which is important for user satisfaction and trust while decision-making [25].

Rule-based explanations aim to align with human reasoning by defining model behavior as interpretable IF–THEN statements. Prior work on rule-based and associative classifiers demonstrated that compact rule sets can be both accurate and interpretable [26–28]. More recent approaches, such as LORE [3], generate local explanations by inducing decision rules using a genetic algorithm and a surrogate decision tree. BARBE [6, 29] extends this line of work by learning association rules surrogates from locally sampled data to explain black-box predictions. Despite their rule-based interpretability, existing rule-based explainers are largely *non-interactive*: explanations are produced for a fixed input, and modifying features typically requires resampling neighborhoods and retraining the surrogate. This limitation reduces their effectiveness for exploratory analysis and user-driven what-if reasoning.

A growing research area emphasizes explanations that support user interaction, allowing users to examine different use cases and test model output decisions [18, 30, 31]. Studies show interactive explanations improve understanding, trust, and satisfaction. Yet most interactive XAI systems are based on feature attribution or counterfactual visualizations, lacking rule-based reasoning with strong fidelity guarantees. Interactivity often comes at the expense of faithfulness or interpretability [32, 33].

BARBiE fills these shortcomings by combining association rule based surrogates with real-time interactivity. Unlike feature-importance methods, BARBiE generates explicit rules that are used to calculate the feature-importance scores. In contrast to previous rule-based explainers, BARBiE provides efficient "what-if" analysis. It provides two-stage pruning and better neighborhood generation to reduce the computational cost and improve the performance. It updates applicable rules and feature importance score without retraining. Finally, BARBiE uses a local fidelity gate. Explanations are returned only when the surrogate faithfully is consistent with the black-box prediction. This combination of faithfulness, rule-based reasoning, and interactivity distinguishes BARBiE from existing XAI approaches.

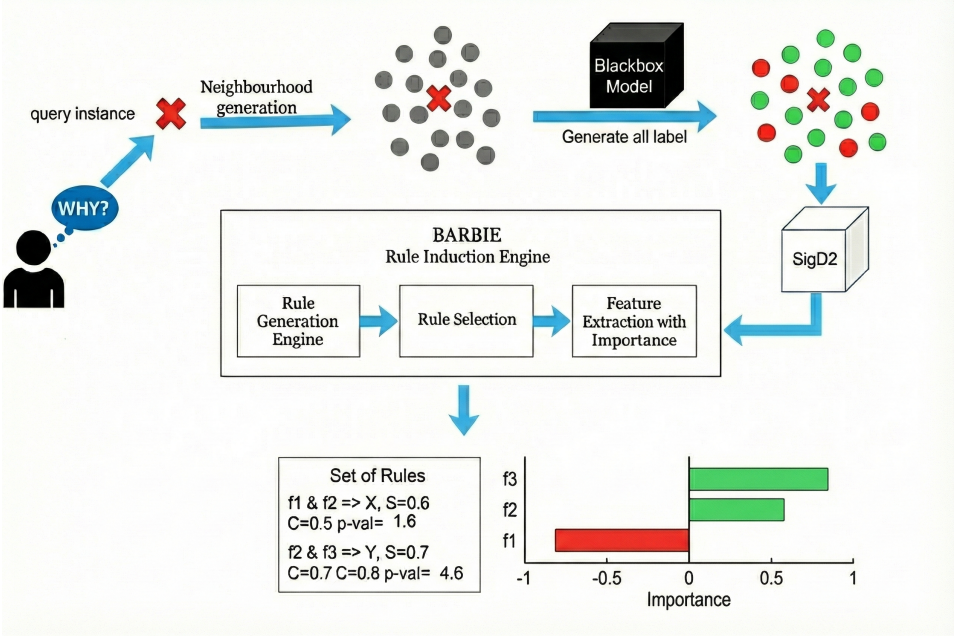


Figure 1. A high level activity diagram of how BARBiE produces explanations for an instance.

3. Methodology

In this section, we describe the **BARBiE** framework for generating faithful, rule-based, and interactive explanations for black-box models. BARBiE uses association rule mining in a surrogate associative classifier model called SigD2 [34] with two stage purning streategy for explaining the reasons behind black-box model outcome decisions. Figure 1 illustrates a high-level activity diagram with a small number of clearly defined steps. Given a trained black-box classifier f and a query instance $\mathbf{x}$, BARBiE generates an explanation by: (i) preparing a stable, rule-friendly representation of the data; (ii) constructing a local neighborhood around $\mathbf{x}$ and querying f for labels; (iii) training a compact, statistically significant rule-based surrogate; (iv) enforcing a local fidelity check; and (v) extracting human-readable rules and rule-grounded feature attributions that update quickly during user interaction.

3.1. 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 a dataset-level preparation phase that is executed once per dataset. The features are partitioned into categorical and numeric types. Later, a lightweight discretization is applied that standardizes both into a rule-friendly space before neighborhood sampling and interpretable learning. During per-instance explanation, categorical features are encoded using a fixed one-hot or ordinal encoding that preserves human-readable semantics. In addition, numeric features are discretized into a small number of interpretable intervals using data-driven binning strategies (e.g., quartiles, deciles, or entropy-based splits). All encoding keys and bin boundaries are frozen at preparation time to ensure stability across explanations and deterministic inverse decoding. This produces a rule-friendly representation while preserving the ability to map rules and perturbations back to the original feature space.

3.2. **Local Neighborhood Generation**

To explain the prediction $f(\mathbf{x})$, BARBiE constructs a synthetic local neighborhood $\mathcal{N}(\mathbf{x}) = \{\mathbf{z}_i\}_{i=1}^m$ centered at 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, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$, where the mean is centered at the query instance and $\mathbf{\Sigma}$ captures feature covariance; alternative heavy-tailed distributions (e.g., t distributions) are optionally supported. Each perturbed point $\mathbf{z}_i$ is inverse-transformed into the original feature space and 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, such as BARBE, which relies on independent normal perturbations, BARBiE samples from a multivariate normal distribution, allowing correlated features to be perturbed jointly. This choice produces a more realistic local neighborhoods and improves surrogate quality, consistent with findings by Smith and Zaïane [35].

3.3. **Interpretable Surrogate Learning**

BARBiE trains a local, interpretable surrogate on $\mathcal{N}(\mathbf{x})$ using SIGD2 [34], a statistically significant associative classifier with two-stage pruning. The surrogate induces short IF–THEN rules of the form $r : A \Rightarrow y$, where A is a conjunction of feature predicates, and y a class label. Given the synthetic neighborhood $\mathcal{N}(\mathbf{x}) = \{(\mathbf{z}_i,\ f(\mathbf{z}_i))\}_{i=1}^m$, SIGD2 is trained to map $\mathbf{z} \mapsto f(\mathbf{z})$ *within* $\mathcal{N}(\mathbf{x})$. Since the target is local rather than global, the surrogate can remain simple while still achieving a high agreement with f around $\mathbf{x}$.

The surrogate induces short conjunctions of literals,

$$r : \bigwedge_{j \in J_r} \pi_j \ \Rightarrow\ y, \qquad |r| = |J_r|,$$

where each predicate π_j corresponds to a human-readable condition (e.g., a categorical value or a numeric interval from a quantile bin). For each rule r we retain: (i) *support* $\mathrm{supp}(r)$, (ii) *confidence* $\mathrm{conf}(r)$, and (iii) a bounded significance score derived from $-\log p(r)$ (capped to avoid domination by extremely rare patterns). Conservative thresholds for support, confidence, significance, and rule length encourage parsimony. Each rule is provided with support, confidence, and a capped statistical significance score derived from Fisher's exact test. In contrast to generic trees or linear surrogates, SIGD2 with two-stage pruning removes redundant, noisy, and low-utility rules, yielding a compact and readable rule set that approximates the behavior of f in the vicinity of $\mathbf{x}$.

3.4. **Local Fidelity Gate**

To ensure 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 $\mathbf{x}$. BARBiE resamples the local neighborhood and retrains the surrogate up to a bounded number of attempts, If the check fails. 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.

3.5. **Rule Extraction**

Once fidelity is confirmed, BARBiE extracts two complementary rule views: (i) the full set of local rules for contextual understanding, and (ii) the subset of rules applicable to $\mathbf{x}$. Rules are decoded into human-readable predicates and ranked to prioritize those aligned with $f(\mathbf{x})$, followed by confidence, support, and statistical significance. In addition, BARBiE retrieves contrastive rules associated with alternative classes to highlight how structured changes in inputs could alter the prediction.

3.6. **Feature Importance**

BARBiE derives per-instance feature importance directly from the surrogate's rule base. For each feature, BARBiE aggregates signed evidence from all rules that mention the feature, weighting each contribution by rule quality (support, confidence, significance) and by the degree to which the rule is satisfied by $\mathbf{x}$. We convert rules into signed per-feature attributions by aggregating rule evidence for every feature j. For each rule r containing j, we define a dynamic satisfaction weight:

$$w_{\text{dyn}}(\mathbf{x}, r) \;=\; \frac{2\, n_{\text{sat}}(\mathbf{x}, r) + 1}{2\,|r| + 2} \;\in\; \left[\frac{1}{2|r|+2},\, 1\right). \tag{3.1}$$

Rule strength combines frequency, precision, and (optional) significance:

$$s(r) \;=\; \text{supp}(r) \cdot \text{conf}(r) \cdot \phi(r), \tag{3.2}$$

$$\phi(r) \in \left\{1,\; \frac{\min\big(-\log p(r),\, L\big)}{L}\right\},\; L > 0. \tag{3.3}$$

Rules that $\mathbf{x}$ satisfies more completely receive a larger weight (a smooth applicability factor), while rules pointing against the predicted class contribute negative evidence. This yields *signed* contributions that capture support *for* vs. *against* $f(\mathbf{x})$. Let $\text{sgn}(r, \hat{y}) \in \{-1, +1\}$ denote whether the rule's consequent supports $(+1)$ or opposes (-1) $\hat{y}$. The raw contribution of r to feature j is

$$\Delta_j(r, \mathbf{x}) \;=\; \text{sgn}(r, \hat{y}) \cdot s(r) \cdot w_{\text{dyn}}(\mathbf{x}, r). \tag{3.4}$$

Summing across rules that involve j gives the unnormalized score

$$\tilde{\omega}_j(\mathbf{x}) \;=\; \sum_{r\,:\,j \in r} \Delta_j(r, \mathbf{x}). \tag{3.5}$$

Finally, we apply ℓ_1-normalization to yield signed importances bounded by ± 1:

$$\omega_j(\mathbf{x}) \;=\; \begin{cases} \dfrac{\tilde{\omega}_j(\mathbf{x})}{\sum_k |\tilde{\omega}_k(\mathbf{x})|}, & \text{if } \sum_k |\tilde{\omega}_k(\mathbf{x})| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

A feature is important if it appears in confident, well-supported, significant rules that the instance largely satisfies and that point toward the predicted class. The feature importance list highlights dominant drivers (large positive scores) and opposing factors (negative scores). Internally, BARBiE also considers the *next-best* local class when summarizing feature usage, which helps disentangle features that chiefly separate the predicted class from its most competitive alternative in $\mathcal{N}(\mathbf{x})$.

4. **Results and Discussion**

We evaluate BARBiE on four widely used tabular datasets containing numeric and categorical attributes: the Loan Approval dataset [36], and the Iris, Glass, and Wine datasets from the UCI Machine Learning Repository [37]. We use Decision Trees (DTs) [38] as the

primary black-box model because they provide (i) explicit, symbolic decision paths that enable transparent inspection of rule extraction, (ii) deterministic and reproducible predictions, and (iii) direct control of decision complexity via a maximum depth parameter. These properties make DTs well suited for evaluating rule-based explanations while remaining independent of the explainer itself.

Our evaluation focuses on two complementary aspects: (i) the *practical usability* of BARBiE for exploratory analysis through interactive *what-if* reasoning, and (ii) *human-centered effectiveness*, measured via a controlled user study assessing informativeness, understandability, trustworthiness, and satisfaction.

4.1. **Qualitative Analysis: Interactive What-If Explanations**

BARBiE supports interactive *what-if* analysis by updating rule applicability and rule-grounded 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 2 illustrates a representative example from the Loan Approval dataset. For a given applicant instance initially predicted as *Declined*, BARBiE presents a concise set of IF–THEN rules identifying credit history as the dominant condition driving the decision. When the user modifies the `Credit_History` feature to indicate that guidelines are met, BARBiE immediately updates the explanation: the original rule is no longer applicable, and a new rule supporting an *Approved* prediction is highlighted (Figure 3).

This interaction demonstrates how BARBiE exposes actionable conditions that govern prediction changes and allows users to reason directly about the model's decision logic. Unlike static explainers, BARBiE enables real-time scenario testing, helping users understand not only *why* a decision was made, but also *how* it could change under plausible input modifications.

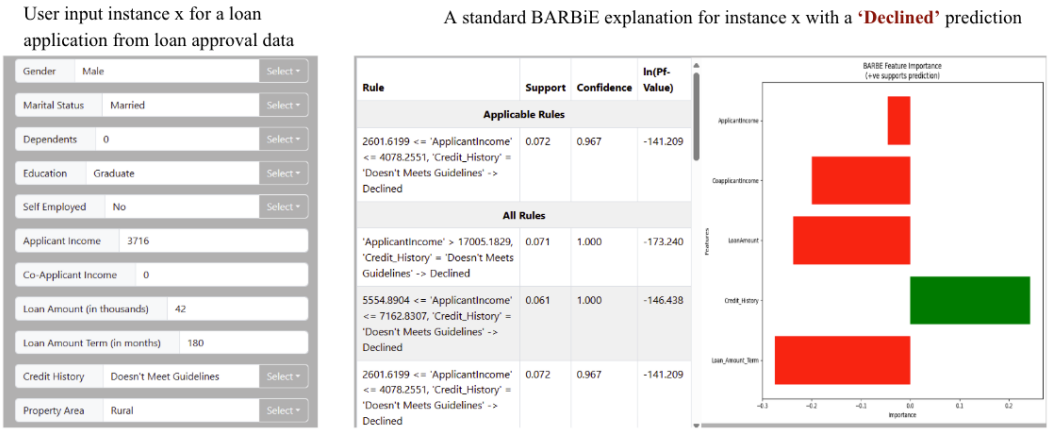


Figure 2. A BARBiE explanation for an instance with a *Declined* prediction, showing the dominant rule conditions.

4.2. **User Study Results**

To assess human-centered outcomes, we conducted a controlled user study following the principles of the DARPA Explainable Artificial Intelligence program [8]. The study protocol was approved by our institution's Research Ethics Board. A total of 34 participants completed the study, with ages ranging from 18 to 50 (mean = 26.02, SD = 6.42); 46%

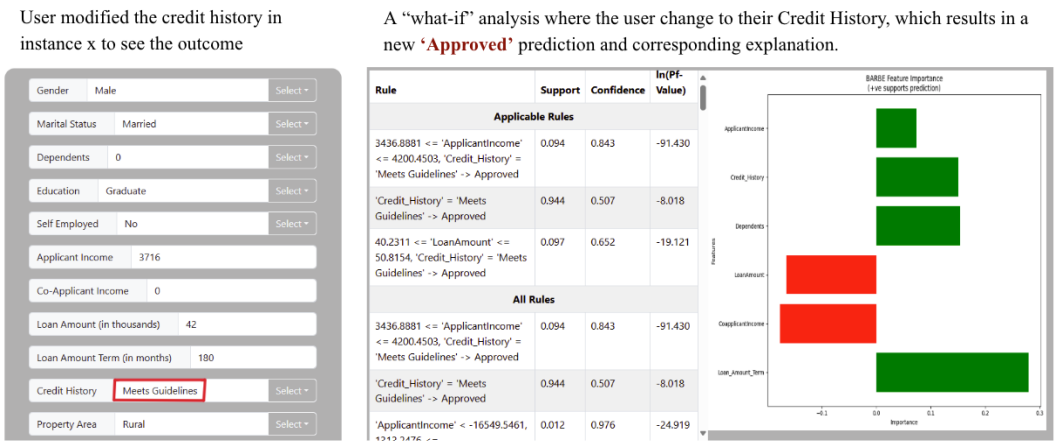


Figure 3. A *what-if* analysis where modifying `Credit_History` results in an *Approved* prediction and an updated rule-based explanation.

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

Method	Info	Understand	Trust	Satisfaction
LIME	3.6±0.7	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

identified as female. Participants reported mixed familiarity with machine learning (60% "very familiar", 40% "somewhat familiar" or "neutral").

Participants compared BARBiE against LIME, SHAP, and BARBE using anonymous, within-task evaluations. Explanations were rated on a 5-point Likert scale across four dimensions: *Informativeness*, *Understandability*, *Trustworthiness*, and *Satisfaction*.

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 feature-importance-based explainers and static rule-based baselines. These results suggest that presenting concise IF–THEN rules together with interactive *what-if* capabilities helps users form clearer mental models of black-box behavior and supports more effective exploration and decision-making.

Overall, the results demonstrate that BARBiE's interactive, rule-based design offers a practical and human-centered alternative to static explanation methods, enabling users to engage with black-box models in a more transparent and actionable manner.

4.3. **Discussion: Interactivity and Improved Rule Surrogates**

BARBiE improves upon prior rule-based explainers, particularly LORE and BARBE, along two key dimensions: *interactivity* and the *quality of the rule-based surrogate*. While LORE use genetic algorithm and BARBE introduced statistically grounded association rules as local surrogates for black-box explanations, their explanations remain static and tightly coupled to a fixed input instance. Any change to feature values typically requires regenerating neighborhoods and retraining the surrogate, which limits their usefulness for exploratory analysis and interactive reasoning. In addition, LORE does not provide any support, confidence and p-value.

In contrast, BARBiE is explicitly designed for interactive *what-if* analysis. Once a local surrogate is trained, users can modify input features and immediately observe updated explanations through changes in applicable rules and rule-grounded feature importance. This design enables rapid scenario testing without retraining the black-box model or re-mining rules, supporting a more natural, iterative form of human reasoning that aligns with real-world decision-making workflows.

Beyond interactivity, BARBiE adopts an improved rule surrogate based on SigD2 with a two-stage pruning strategy. Compared to the original SigDirect-based surrogate used in BARBE, SigD2 produces more concise and readable rule sets by systematically removing redundant, low-utility, and overly specific rules. This pruning process reduces cognitive load for users and improves the clarity of explanations by prioritizing short, high-quality IF–THEN rules that capture salient feature interactions. As a result, explanations presented by BARBiE are not only interactive but also more interpretable and actionable.

Together, these design choices distinguish BARBiE from BARBE and other static rule-based explainers. By combining an improved, pruned rule surrogate with real-time interaction, BARBiE shifts the role of explanations from passive artifacts to active tools for understanding and exploring black box model behavior.

5. Conclusion and Future Work

This paper introduced **BARBiE**, a model-agnostic framework for explaining black-box predictions on tabular data through interactive, rule-based local explanations. BARBiE combines an improved association-rule surrogate based on SigD2 with a two-stage pruning strategy, a lightweight instance-level fidelity gate, and rule-grounded feature importance to produce compact and interpretable IF–THEN explanations aligned with human rule-oriented reasoning.

Unlike static feature-attribution methods, BARBiE enables real-time *what-if* analysis by updating rule applicability and feature importance as users modify inputs without the need to retrain the surrogate or the black-box model. This design shifts explanations from passive summaries to interactive tools that support exploratory reasoning and scenario testing.

Through experiments on four popular datasets and a controlled user study, BARBiE consistently received higher ratings for informativeness, understandability, trust, and satisfaction compared to LIME, SHAP, and BARBE. The findings indicate that interactive, rule-based explanations are more informative, satisfactory, trustworthy, and understandable than static attribution-based approaches. Because BARBiE operates without access to model internals and relies only on prediction APIs, it is well-suited for realistic enterprise and institutional deployment settings.

In future we will focus on improving scalability through more aggressive rules pruning, supporting explicit counterfactual rule generation, and advancing the framework beyond tabular data to text and vision domains.

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] 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).

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

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

[6] 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.

[7] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

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

[9] S. Kabir, M. S. Hossain, and K. Andersson. "A review of explainable artificial intelligence from the perspectives of challenges and opportunities". In: *Algorithms* 18.9 (2025), p. 556.

[10] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". In: *Information fusion* 99 (2023), p. 101805.

[11] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv preprint arXiv:1312.6034* (2013). URL: https://arxiv.org/abs/1312.6034.

[12] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74. URL: https://doi.org/10.1109/ICCV.2017.74.

[14] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. "Sanity checks for saliency maps". In: *Advances in neural information processing systems* 31 (2018).

[15] S. Jain and B. C. Wallace. "Attention is not Explanation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Association for Computational Linguistics, 2019, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://doi.org/10.18653/v1/N19-1357.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin. "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386* (2016).

[17] M. T. Ribeiro, S. Singh, and 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)*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

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

[19] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz. "A perspective on explainable artificial intelligence methods: SHAP and LIME". In: *Advanced Intelligent Systems* 7.1 (2025), p. 2400304.

[20] A. B. Haque, A. N. Islam, and P. Mikalef. "Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research". In: *Technological Forecasting and Social Change* 186 (2023), p. 122120.

[21] D. Alvarez-Melis and T. S. Jaakkola. "On the robustness of interpretability methods". In: *arXiv preprint arXiv:1806.08049* (2018).

[22] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo. "Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models". In: *Journal of the Operational Research Society* 73.1 (2022), pp. 91–101.

[23] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems* 30 (2017).

[24] I. E. Kumar et al. "Shapley Residuals: Quantifying the Limits of the Shapley Value for Explanations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[25] M. Guesmi, M. A. Chatti, J. Ghorbani-Bavani, S. A. Joarder, Q. U. Ain, and R. Alatrash. "What if Interactive Explanation in a Scientific Literature Recommender System." In: *IntRS@ RecSys*. 2022, pp. 104–121.

[26] 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.

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

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

[29] M. H. Motallebi Shabestari. "Explaining Decisions of Black-box Models using Association Rules". In: (2020).

[30] R. Binns. "What can political philosophy teach us about algorithmic fairness?" In: *IEEE Security & Privacy* 16.3 (2018), pp. 73–80.

[31] 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.

[32] C. Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[33] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 2020, pp. 180–186. DOI: 10.1145/3375627.3375830.

[34] 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.

[35] 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).

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

[37] 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.

[38] J. Su and H. Zhang. "A fast decision tree learning algorithm". In: *Aaai*. Vol. 6. 2006, pp. 500–505.