Quantifying and Mitigating Epistemic Uncertainty in Local Rule-Based Explanations

Szymon Bobek[1][0000−0002−6350−8405] and
Maciej Mozolewski[1][0000−0003−4227−3894]

Department of Human-Centered Artificial Intelligence, Institute of Applied Computer Science, Jagiellonian University, Łojasiewicza 11, Krakow, 30-348, Poland
{szymon.bobek, m.mozolewski}@uj.edu.pl

Abstract. Uncertainty is intrinsic to statistical learning, arising from multiple sources. One recently examined form is epistemic uncertainty, which stems from the difficulty humans face in understanding or inspecting the internal workings of black-box models. Explainable AI (XAI) aims to mitigate this by revealing how models operate. However, local post-hoc explainers can sometimes have the opposite effect. In particular, local rule-based methods may produce contradictory explanations across different neighborhoods, undermining user trust and obscuring the model's global logic. We propose a framework for detecting, quantifying, and mitigating inconsistencies among local rule-based explanations. Our contributions include Conflict-Conditioned Empirical Disagreement under Uncertainty (CC-EDU), a metric for neighborhood-level inconsistency and a restriction mechanism that refines overly general rules. Experiments on selected benchmark datasets show that our framework correctly detects and reduces inter-rule contradictions while preserving fidelity.

Keywords: Epistemic Uncertainty · Uncertainty Quantification · Rule-Based Systems · Explainable AI

1 Introduction

Uncertainty is unavoidable in modern computational science applications, ranging from medical diagnosis and industrial predictive maintenance to environmental modelling and economic forecasting. In such domains, decisions increasingly rely on machine learning models trained on large-scale, heterogeneous, and potentially noisy data, introducing different types of uncertainty. Typically, we distinguish between aleatoric uncertainty, which stems from inherent noise or randomness in the data, and epistemic uncertainty, which arises from limited knowledge, insufficient data, or model inadequacies.

In practice, explainable AI (XAI) methods are often used to reduce epistemic uncertainty regarding model behavior by giving insight into how model derives its decision. Among different explanation types, such as feature-attribution explanations [17, 22, 26, 25], counterfactual explanations [13], visual explanations [12]

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

and rule-based ones [14, 21, 8, 18, 23], empirical studies indicate that experts and users strongly prefer those that offer structural clarity, explicit decision logic, and alignment with human reasoning patterns [5, 4].Rules naturally reflect decision boundaries in terms of constraints on features and are widely used in professional reasoning in medicine, engineering, and finance.

Most rule-based explainers operate locally. They approximate the decision boundary of a predictive model in a neighborhood around a selected instance. Methods such as LUX [9], LORE [14], and EXPLAN [21] generate synthetic neighborhoods and train shallow decision trees. Other approaches prioritize local fidelity through different mechanisms, such as Anchors [23] which relies on constrained optimization, RuleXAI [18] which leverages rule induction, and PHAR [19] which employs local perturbations to extract interpretable interval bounds from numeric feature attributions such as SHAP.

However, the reliance on locality introduces an additional source of epistemic uncertainty that arises from the modeling and approximation choices involved in constructing local explanations. In particular, the definition of the neighborhood around a query instance – including its size, geometry, and sampling strategy – directly influences the surrogate rule that is extracted. Different yet equally plausible neighborhood specifications may therefore yield different local rules for the same instance. Furthermore, when neighborhoods overlap across nearby instances, the resulting locally faithful rules may remain individually valid while becoming mutually contradictory as depicted in Figure 1. This variability reflects uncertainty stemming from limited knowledge about the true local decision boundary and the methodological assumptions used to approximate it.

A similar situation may arise in medical decision-making, or industrial applications where a locally accurate rule may contradict established domain gold standards introducing uncertainty and doubt among experts interpreting the explanation. In this work, we address this problem by introducing a framework for detecting and resolving contradictions among local rule-based explanations. We argue that explanation inconsistency itself constitutes a form of uncertainty about which explanation is correct that must be quantified and mitigated.

2 Related Work

Uncertainty quantification in computational science has traditionally centered on data and model uncertainty, but recent XAI research shows that uncertainty also arises within explanations, through instability and disagreement among different methods. While prediction-level uncertainty is well characterized, explanation-level uncertainty remains comparatively underexplored [10]. One direction of research in this area formalizes the disagreement problem, showing that widely used post-hoc explainers such as LIME [22], SHAP [17], or Integrated Gradients [26] can yield very different attribution patterns for the same prediction. This observation has motivated consensus-oriented approaches, including explainer-agreement regularization to encourage alignment between attribution methods [24], minimizing uncertainty by constructing weighted ensemble ex-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

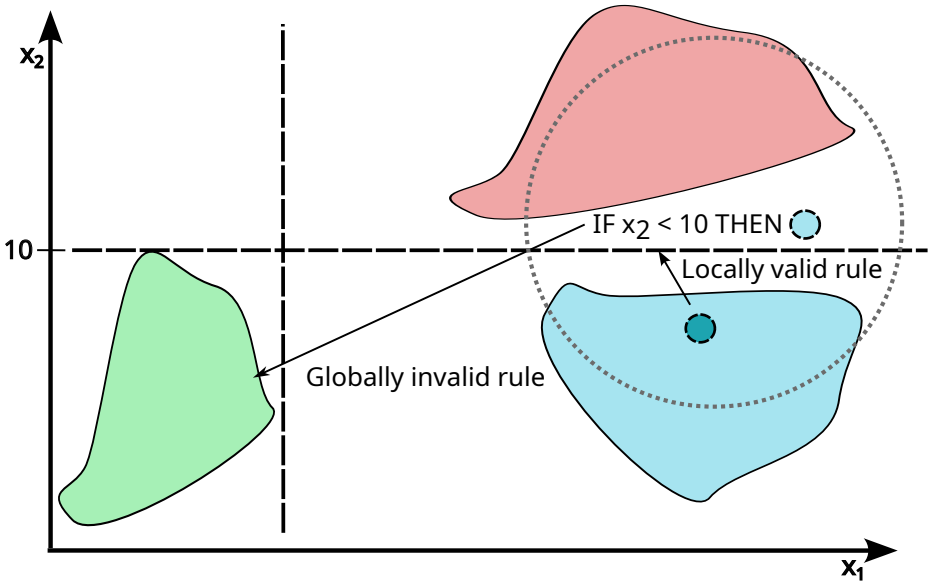

Fig. 1. Epistemic uncertainty induced by locality in post-hoc local explanation methods. Variations in the definition of the neighborhood around a query instance—such as its size, geometric shape, or sampling strategy—can lead to different surrogate rules for the same data point, revealing locality-driven instability in the explanation.

plainers [3] as well as quantitative frameworks that systematically characterize disagreement in explainable machine learning [15]. In the domain of rule-based explanations, overlapping and contradictory outputs are often managed through conflict resolution strategies (e.g., [19]). These inconsistencies frequently stem from the Rashomon effect [20, 16]-a phenomenon observed in both predictive models and their explanations, where multiple distinct, yet equally faithful representations exist for the same data.

Other contributions show that explanations frequently fail to reliably propagate uncertainty under perturbations [11]. In our earlier work, we also examined how uncertainty present in model predictions should be faithfully conveyed through explanatory outputs, emphasizing that explanation reliability depends on the propagation of predictive uncertainty into the explanation space [7].

However, despite these advances, most existing efforts remain focused on feature-attribution methods or uncertainty propagation from model to explainer. As a result, these approaches often fail to capture uncertainty inherent in rule-based explanations, which becomes evident when local rules—derived for individual instances—are applied across broader neighborhoods or the entire dataset, revealing overlapping decision regions that can produce logically inconsistent or contradictory predictions.

Rule-based explanations uniquely provide human-readable decision logic, explicitly define regions of the input space, and are widely used in local surrogate

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

methods because of their interpretability and alignment with expert reasoning. These properties make them particularly suitable for analyzing structural inconsistencies, logical contradictions, and local-global conflicts–yet existing XAI uncertainty frameworks do not quantify or mitigate inter-rule contradictions. Accordingly, our work addresses uncertainty at the level of explanations themselves by introducing CC-EDU, a neighborhood-conditioned metric that directly measures empirical disagreement among overlapping local rules, together with a conflict-aware refinement strategy operating in the explanation space rather than on model parameters. This fills a crucial gap–moving from merely quantifying uncertainty in explanation outputs to actively mitigating structural inconsistency in rule-based XAI.

3 Method

In this section, we present a formal framework for detecting, quantifying, and mitigating contradictions among local rule-based explanations. Our approach operates directly in the explanation space and evaluates conflicts empirically within instance-specific neighborhoods. This allows us to reduce epistemic uncertainty introduced by overlapping local explanations while preserving local fidelity.

3.1 Problem Formulation

Let $f : \mathbb{R}^d \to \mathcal{Y}$ denote a predictive model over a d-dimensional feature space with output space $\mathcal{Y}$. Let $X = \{x_i\}_{i=1}^n$ be a dataset of n instances. A local rule-based explainer produces a set of rules $\{R_i\}_{i=1}^n$, one for each instance x_i. Each rule R_i is defined as:

$$R_i : \quad x \in C_i \implies \hat{y}_i,$$

where C_i denotes the subset of instances empirically covered by the rule's conditions, and $\hat{y}_i \in \mathcal{Y}$ is the predicted class assigned by the rule.

Traditionally, a rule is represented intensionally as a conjunction of feature-level constraints. Such a representation induces a decision region in the input space, defined as the set of points satisfying those constraints. When restricted to the empirical dataset X, this region corresponds exactly to the coverage set C_i. In this work, we adopt this extensional perspective and identify each rule with its induced decision region over the data, represented by C_i.

This formulation is intentional. Since our objective is to quantify empirical overlap, contradiction, and disagreement among local rules, the relevant object of analysis is the subset of data points jointly covered by different rules. Representing rules via their coverage sets therefore allows us to define and measure empirical disagreement directly in terms of set intersections, without dependence on the specific syntactic form of the rule conditions. By evaluating intersections based on actual data samples rather than analytical geometric volumes, the method naturally accounts for the underlying density of the data distribution. This prevents the detection of phantom conflicts in empty regions of the feature space and focuses the refinement process strictly on overlapping areas that are empirically supported by observations.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

Definition 1 (Neighborhood-Conditioned Rule Contradiction)

Two rules R_i and R_j are said to contradict each other *with respect to instance* x_i if they predict different classes and overlap sufficiently within the neighborhood of x_i as shown in Equation (1).

$$\text{contr}_i(R_i, R_j) = \begin{cases} 1 & \text{if } \hat{y}_i \neq \hat{y}_j \text{ and } \frac{|(C_i \cap C_j) \cap \mathcal{N}_k(i)|}{|(C_i \cup C_j) \cap \mathcal{N}_k(i)|} \geq \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here $|\cdot|$ denotes empirical cardinality over dataset indices, $\tau \in [0, 1]$ is a minimum neighborhood overlap threshold and $\mathcal{N}_k(i) \subseteq \{1, \ldots, n\}$ denote the indices of k nearest neighbors of instance x_i in feature space.

Definition 2 (Rule Structural Disagreement)

The Structural Disagreement between two rules quantifies the degree of agreement defined as a weighted combination of empirical coverage overlap $\frac{|C_i \cap C_j|}{|C_i \cup C_j|}$, shared feature-condition overlap F_{o}, and confidence penalty Conf_{p} when $\hat{y}_i = \hat{y}_j$ as defined in Equation (2). The intuition behind this is that disagreement should be strongest when two rules not only collide in terms of empirical coverage but are also structurally similar—using the same features and making confident predictions. The constant 0.5 serves as a baseline weight ensuring that spatial overlap remains the dominant component of the score, since the multiplicative factor lies within $[0.5, 1]$. The additional terms refine this baseline: the feature-overlap term F_{o} measures how similar the two rules are in terms of the features they condition on, while the confidence penalty reflects the degree of misalignment in the rules' predictive confidence: the penalty increases when the rules' predictive confidences diverge, and it decreases when a rule's confidence is low, reducing its influence on the overall disagreement. These terms modulate the score without overwhelming the contribution of empirical co-coverage, and the baseline 0.5 prevents the similarity from collapsing to zero for rules that fully overlap in space but differ in structure.

$$\text{disagree}(R_i, R_j) = \frac{|(C_i \cap C_j) \cap \mathcal{N}_k(i)|}{|(C_i \cup C_j) \cap \mathcal{N}_k(i)|} \cdot (0.5 + \alpha F_{\text{o}}(R_i, R_j) + \beta \text{Conf}_{\text{p}}) \tag{2}$$

The parameters α and β can be adjusted depending on the desired sensitivity of each component, with default values set to 0.3 and 0.2, respectively. This setting establishes a clear hierarchy of importance within the metric, ensuring that empirical spatial overlap remains the dominant factor, followed by the structural similarity of the features utilised, while the predictive confidence penalty serves as the least influential modifier.

Conflict-Conditioned Empirical Disagreement (CC-EDU)

The CC-EDU score for rule R_i is defined as:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

$$\text{CC-EDU}(R_i) = \frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k(i)} \text{contr}_i(R_i, R_j) \cdot \text{disagree}(R_i, R_j). \tag{3}$$

High CC-EDU indicates strong empirical disagreement among locally overlapping explanations. Since contradiction is explicitly conditioned on $\mathcal{N}_k(i)$, CC-EDU measures *instance-specific epistemic inconsistency*.

The metric captures situations where multiple faithful rules coexist in the same region of the feature space yet induce conflicting predictions. Such conflicts may not significantly affect aggregate predictive performance, but they reveal instability in the explanation space. Consequently, CC-EDU serves as an indicator of explanation robustness, identifying regions where interpretability is epistemically uncertain despite stable model outputs. Based on this metric a set of rules for conflict-aware restriction is selected.

Computational Complexity

Empirical neighborhood-based overlap analysis determines the baseline time complexity of the proposed framework. Unlike the subsequent conflict-aware restriction, which is executed conditionally only when disagreement thresholds are exceeded, detecting contradictions requires a systematic evaluation of rule intersections across the data. Evaluating the CC-EDU metric globally incurs a cost directly tied to its empirical nature. Specifically, the algorithm iterates over all n instances, pairing each rule with the k rules originating from its local neighborhood. For every pair, assessing the empirical intersection requires evaluating the rule conditions on k local data points across the d-dimensional feature space, yielding a cost of $\mathcal{O}(k \cdot d)$ per comparison. Consequently, the cross-evaluation of k pairs for all n instances results in an overall computational complexity of $\mathcal{O}(nk^2d)$ for the contradiction detection phase. In the limiting case where the neighborhood encompasses the entire dataset ($k = n$), this worst-case complexity simplifies to $\mathcal{O}(n^3d)$.

Conflict-Aware Rule Restriction

To reduce local contradictions, overly general rules are refined. Let R_i be in contradiction with R_j under Definition 1. A restricted rule is defined as:

$$R_i^{\text{restricted}} : x \in C_i' \implies \hat{y}_i,$$

where $C_i' \subseteq C_i$ is obtained by adding additional feature constraints derived from the empirical disagreement region

$$D_{ij} = (C_i \cap C_j) \cap \mathcal{N}_k(i).$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

Restriction is applied only if $x_i \in C_i'$ (the explained instance remains covered), and the directional disagreement strength exceeds a predefined threshold δ as given in Equation (4).

$$\frac{|D_{ij}|}{|C_i \cap \mathcal{N}_k(i)|} \geq \delta \tag{4}$$

The additional restriction condition is derived directly from the empirical disagreement region D_{ij}. In practice, we construct a simple local decision stump trained only on samples within D_{ij}, where the two rules disagree. This stump identifies the most discriminative feature and threshold separating conflicting predictions, hence providing a minimal additional constraint. The resulting split is then incorporated into overgeneral rule, yielding the refined condition $C_i' \subseteq C_i$. Importantly, the restriction is purely local and data-driven: it does not alter the rule outside the neighborhood nor introduce external assumptions, but instead sharpens the rule exactly in the region where contradiction is observed. Since restriction enforces $C_i' \subseteq C_i$, neighborhood-conditioned overlap can only decrease or remain unchanged, ensuring that CC-EDU cannot increase under rule shrinking.

Locality Selection

The neighborhood size k determines the scale at which empirical contradiction is evaluated. For each instance x_i, contradictions are assessed within its k-nearest neighborhood $\mathcal{N}_k(i)$. Thus, k controls the sensitivity of disagreement detection: small values of k focus on highly local inconsistencies, while larger values progressively approximate global contradiction. In the limiting case $k = n$, contradiction is evaluated at the dataset level.

Importantly, k is not optimized in our framework. It is a user-specified parameter reflecting the scale of explanation stability deemed relevant by the expert. Different choices of k correspond to different operational definitions of locality, and therefore to different notions of epistemic sensitivity.

In the empirical study presented later, we report results across a range of k values to illustrate the robustness of the proposed refinement procedure under varying locality assumptions.

4 Evaluation

4.1 Experimental Setup

We demonstrated the proposed conflict-aware refinement procedure on several standard benchmark datasets, including Breast Cancer, Iris, Wine, and Pima from the UCI Repository [1] and OpenML [2] Credit-G and Sonar. These datasets differ in dimensionality, class distribution, and boundary complexity, providing heterogeneous conditions for assessing both predictive fidelity and empirical rule

[1] See: https://archive.ics.uci.edu/ml/index.php.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

disagreement. We used LUX as our explainer, as it has demonstrated to deliver the highest quality explanations among state of the art rule-based explainers for tabular data [8].

For each dataset, a local rule-based explainer (LUX) is first applied to generate one rule per instance at a given locality level. We then compute the average CC-EDU score, which quantifies empirical contradiction among overlapping local rules. Next, the proposed conflict-aware restriction mechanism is applied, and both CC-EDU and predictive fidelity are recomputed.

To ensure robustness, the evaluation is conducted over a grid of locality hyperparameters. The neighborhood size k varies from 5 to 100 (in steps of 5, bounded by dataset size), while the locality scaling parameter λ ranges from 0.01 to 0.2. This results in up to 300 configurations per dataset. All configurations are evaluated independently and averaged to obtain dataset-level summaries.

In our experiments, both hyperparameters related to our approach – the overlap threshold τ from Equation (1) and the restriction threshold δ introduced in Equation (4) – are fixed at 0.01. This means that two rules are considered contradictory if their coverage overlaps by at least 1% within the neighborhood of an instance, and a rule is refined only if at least 1% of its covered neighborhood is involved in such a contradiction. Setting these low, conservative values ensures that even small overlaps and minor contradictions are detected and addressed, while avoiding overly aggressive pruning of rules. Maintaining equality between these two thresholds ($\tau = \delta$) establishes a uniform proportional strictness for conflict management. Thus we demand the same relative evidence: a 1% overlap against the joint coverage to flag a contradiction, and a 1% overlap against the rule's individual coverage to trigger its structural refinement.

4.2 Metrics

Predictive fidelity is measured using the micro-averaged F1 score, computed between the rule-based predictions and the ground-truth labels. Because local explainers typically produce precise rule-based explanations—and therefore each rule evaluated on the instance it was generated for yields an F1 score of 100% – we compute the overall F1 score across all instances and rules using a weighted – voting scheme, where the weights correspond to the rules' confidence values. Let $F1_{\text{base}}$ denote the baseline fidelity and $F1_{\text{r}}$ the fidelity after restriction. We report the relative change in fidelity as:

$$\text{F1 Gain} = \frac{\mathbb{E}[F1_{\text{r}} - F1_{\text{base}}]}{\mathbb{E}[F1_{\text{base}}]}. \tag{5}$$

Empirical disagreement is measured using CC-EDU, where lower values indicate less contradiction among overlapping rules. Let EDU_{base} and EDU_{r} denote the values before and after refinement, respectively. The relative contradiction reduction is defined as:

$$\text{EDU Drop} = \frac{\mathbb{E}[EDU_{\text{base}} - EDU_{\text{r}}]}{\mathbb{E}[EDU_{\text{base}}]}. \tag{6}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

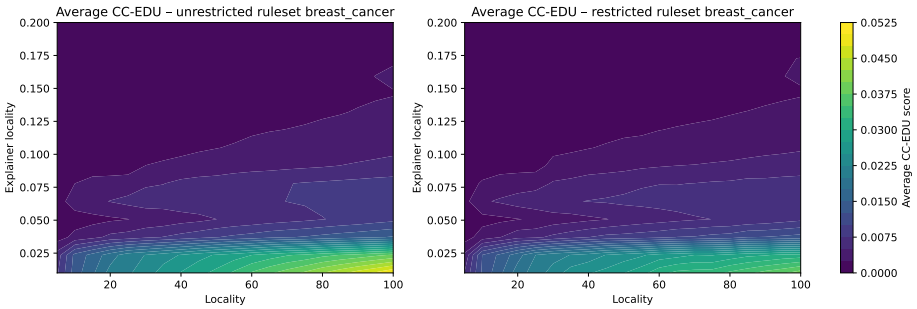

Fig. 2. Heatmap of CC-EDU across varying neighborhood size k and LUX locality settings for the Breast Cancer dataset.

All reported values correspond to averages over valid hyperparameter configurations.

4.3 Results

Across 1350 evaluated configurations, the proposed refinement consistently reduces empirical contradiction. The macro-average CC-EDU reduction across datasets equals 10.58%, with dataset-specific reductions ranging from 5.68% (Sonar) to 13.47% (Iris). Most datasets exhibit reductions between roughly 9% and 13%, indicating a stable and systematic decrease in overlapping contradictory regions in the explanation space.

Figure 2 shows the average CC-EDU values across combinations of neighborhood size k and LUX locality parameters. Lower values indicate reduced empirical contradiction. The consistent decrease across locality scales demonstrates that the restriction mechanism robustly suppresses overlapping contradictory rule regions, largely independent of the chosen locality configuration.

In contrast, predictive fidelity remains essentially unchanged and shows a slight overall improvement. The macro-average relative change in F1 equals +0.12%, indicating a negligible but positive aggregate effect. While two datasets exhibit small decreases in fidelity (Breast Cancer: -0.65%, Pima: -0.48%), the remaining datasets show minor improvements. All changes remain well below one percentage point in magnitude, confirming that contradiction reduction is achieved without materially affecting predictive performance.

Figure 3 presents how varying the neighborhood size k and the LUX locality setting affects F1 performance, highlighting the contrast between invariant unrestricted rules and the k-sensitive behavior of restricted rules.

Under unrestricted rules, F1 remains nearly constant across k because predictions are aggregated via weighted voting. Even if larger neighborhoods introduce contradictory rules, those closer to the instance still dominate the vote, keeping the final classification stable.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

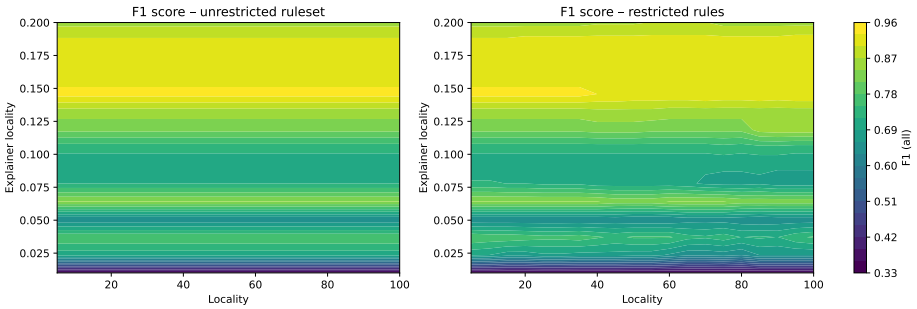

Fig. 3. Heatmap of F1 scores across varying neighborhood size k and LUX locality settings for the Breast Cancer dataset.

Under restricted rules, overlapping coverage decreases, so contradictions are no longer smoothed out by voting. As a result, variations in k produce more visible – yet still minor – perturbations in F1. This further illustrates that F1 alone is insufficient to detect contradictory rule overlap, as reductions in empirical disagreement can occur while predictive fidelity remains practically unchanged. The summary of the results discussed in this section is given in Table 1.

Table 1. Predictive fidelity and empirical disagreement before and after conflict-aware restriction. All values are averaged over evaluated hyperparameter configurations.

Dataset	$F1_{\text{base}}$	$F1_r$	F1 Gain	EDU_{base}	EDU_r	EDU Drop
Breast Cancer	0.7766	0.7715	-0.0065	0.0068	0.0059	0.1301
Iris	0.8756	0.8856	0.0114	0.0218	0.0189	0.1347
Credit-G	0.8053	0.8063	0.0012	0.0250	0.0226	0.0958
Sonar	0.6392	0.6419	0.0043	0.0572	0.0540	0.0568
Pima	0.8462	0.8421	-0.0048	0.0245	0.0217	0.1134
Wine	0.5457	0.5464	0.0014	0.0169	0.0152	0.1039
Macro avg	0.7481	0.7490	0.0012	0.0254	0.0230	0.1058

Column definitions: $F1_{\text{base}}$, $F1_r$: baseline and post-restriction predictive fidelity; F1 Gain: relative F1 gain, see Eq. (5); EDU_{base}, EDU_r: baseline and post-restriction empirical disagreement; EDU Drop: relative reduction in CC-EDU, see Eq. (6).

Importantly, the reduction in contradiction is achieved without degradation in predictive alignment. This suggests that empirical conflicts are concentrated in locally unstable or overlapping regions that contribute disproportionately to disagreement but only marginally to global predictive performance.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

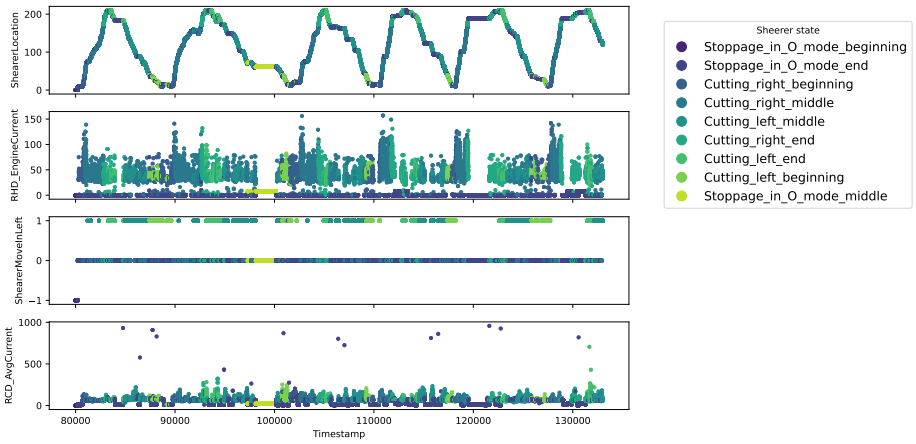

Fig. 4. Sample sensory readings from a coal-mine shearer with operational states marked with colours.

5 Use case

This use case applies CC-EDU to a coal-mine shearer state-classification task with existing, formal expert rules as prior knowledge—a key difference from our earlier work [6], where experts inspected clustering outcomes and refined the knowledge *manually*. Here, CC-EDU automatically detects inconsistencies between locally discovered, rule-based explanations of a classifier and the expert rule base, then guides conflict-aware restriction with CC-EDU to align explanations with domain knowledge.

Accurate and well-explained recognition of shearer operational states underpins safety monitoring and predictive maintenance: operators must not only know *what* state the machine is in, but also *why*, to diagnose anomalies, prevent unsafe behaviour, and plan interventions from sensor evidence. Figure 4 shows sample sensory readings with classification results marked with colours.[2]

We trained a RandomForest classifier on shearer logs, then used LUX to extract a compact set of local, rule-based explanations that cover the prediction space. We injected the expert rule base into the explanation space and treated it as protected: whenever a discovered rule contradicted an expert rule, only the discovered rule could be restricted. CC-EDU was computed in its global setting, aggregating contradictions over the full test distribution, with sensitive overlap/restriction thresholds $\delta = \tau = 10^{-3}$) to capture small but systematic conflicts. We then applied conflict-aware restriction to shrink only those discovered rules implicated in contradictions with expert rules, re-evaluated global CC-EDU, and measured alignment via micro-F1 by executing the rule sets on the test data before and after restriction.

[2] The dataset is publicly available at: https://gitlab.geist.re/pml/x_benchmark-with-selected-datasets.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

In this protected-prior setting, CC-EDU revealed explanation-level contradictions between discovered rules and the expert rule base. After conflict-aware restriction, the mean global CC-EDU decreased from 0.0297 to 0.00921 (CC-EDU Drop = 68.99%) and predictive alignment improved (F1 Gain = 15.31%)[3]. This was in contrary to previously observed negative fluctuations of F1 score reported in evaluation on benchmark datasets in Section 4 and shows that the overall impact of proposed restriction to the quality of the resulting rule-set highly depends on the data.

6 Discussion and Conclusion

The empirical results support the central claim of this work: empirical disagreement among local rule-based explanations can be reduced through conflict-aware restriction while preserving predictive fidelity. Across 1350 evaluated configurations, the proposed refinement achieves a macro-average CC-EDU reduction of 10.58%, while inducing only negligible changes in F1 (macro-average relative change of +0.12%). These results show that the explanation-level inconsistency can be mitigated with almost no cost in fidelity performance. The observed minor fluctuations in F1 are expected. Restricting rules necessarily reduces their coverage in certain neighborhoods, potentially removing both contradictory and correct predictions. However, contradiction tends to concentrate in locally unstable or overlapping regions of the input space. Consequently, refinement primarily eliminates epistemic inconsistency rather than predictive signal. In this sense, the method behaves as a structural regularizer in explanation space: it simplifies overlapping decision regions while preserving their essential predictive behavior.

More broadly, the findings suggest that local faithfulness alone is insufficient to guarantee explanation reliability. Because local post-hoc methods typically omit the implicit neighborhood precondition from their final output, they often project an illusion of global validity. This structural omission can trigger cognitive biases, prompting users to blindly over-generalize local insights to broader regions of the feature space [1]. While ideally an explanation interface should explicitly display both the extracted rule and its defining bounding context to prevent this unbounded extrapolation, visualizing complex high-dimensional neighborhoods for human experts is rarely feasible in practice. Even when individual rules are locally accurate, their mutual overlap may induce contradictions that introduce epistemic uncertainty. Such inconsistency is particularly problematic in high-stakes domains, where explanation stability is as important as predictive performance. By explicitly modeling disagreement through CC-EDU, we extend uncertainty quantification beyond predictive variance and into the structure of explanations themselves.

The proposed framework contributes two key elements. First, CC-EDU provides a principled metric for quantifying empirical contradiction among local rules via their induced coverage sets. Second, conflict-aware restriction offers

[3] For source codes and use case examples see: https://github.com/sbobek/ruledisagree.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

a targeted refinement mechanism that reduces instability without resorting to naive pruning or global majority voting, thereby preserving instance-level interpretability and locality. Additionally, in this work we focused solely on rules derived automatically by the explainer; however, in practice, the explanation rule base can be enriched with expert-defined rules that act as safeguards, against which our framework can detect and resolve contradictions.

Several limitations remain. The current implementation involves non-trivial computational cost due to neighborhood-based overlap analysis, which may require optimization for very large datasets. Additionally, the framework assumes axis-aligned rule structures; extending the approach to oblique or more expressive rule forms constitutes a promising direction for future work. Furthermore, because the contradiction metric relies on empirical coverage, it is inherently sensitive to sampling variance and data drift. Changes in the underlying data distribution may alter the observed rule overlaps, requiring re-evaluation of the explanation consistency in dynamic environments. A comprehensive sensitivity analysis of the introduced hyperparameters, such as the overlap thresholds (τ, δ) and disagreement weights (α, β), was not conducted in this study and remains an important direction for future research to fully understand their impact across diverse data configurations. Finally, we demonstrated the method's performance on one rule per instance scenario, it would be valuable to explore how the framework performs when multiple rules are generated per instance, as this could introduce additional layers of complexity in terms of rule interactions and contradictions.

In summary, we introduced a formal framework for detecting and resolving contradictions among local rule-based explanations. By framing explanation inconsistency as an epistemic uncertainty, we integrate conflict analysis into the broader discourse on uncertainty-aware AI. The results demonstrate that explanation consistency can be improved with minimal predictive cost, highlighting the importance of consistency-aware explanation generation in critical applications.

Acknowledgments

This paper is part of a project that has received funding from the European Union's Horizon Europe Research and Innovation Programme, under Grant Agreement number 101120406. The paper reflects only the authors' view and the EC is not responsible for any use that may be made of the information it contains.

References

1. Astrid Bertrand, Rafik Belloum, James Eagan, and Winston Maxwell. How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91. Association for Computing Machinery, 2022.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

2. Bernd Bischl, Giuseppe Casalicchio, Taniya Das, Matthias Feurer, Sebastian Fischer, Pieter Gijsbers, Subhaditya Mukherjee, Andreas C Müller, László Németh, Luis Oala, Lennart Purucker, Sahithya Ravi, Jan N van Rijn, Prabhant Singh, Joaquin Vanschoren, Jos van der Velde, and Marcel Wever. Openml: Insights from 10 years and more than a thousand papers. *Patterns*, 6(7):101317, 2025.
3. Szymon Bobek, Paweł Bałaga, and Grzegorz J. Nalepa. Towards model-agnostic ensemble explanations. In Maciej Paszynski, Dieter Kranzlmüller, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M.A. Sloot, editors, *Computational Science – ICCS 2021*, pages 39–51, Cham, 2021. Springer International Publishing.
4. Szymon Bobek, Paloma Korycińska, Monika Krakowska, Maciej Mozolewski, Dorota Rak, Magdalena Zych, Magdalena Wójcik, and Grzegorz J. Nalepa. Dataset resulting from the user study on comprehensibility of explainable AI algorithms. *Scientific Data*, 12(1):1000, June 2025.
5. Szymon Bobek, Paloma Korycińska, Monika Krakowska, Maciej Mozolewski, Dorota Rak, Magdalena Zych, Magdalena Wójcik, and Grzegorz J. Nalepa. User-centric evaluation of explainability of ai with and for humans: A comprehensive empirical study. *International Journal of Human-Computer Studies*, 205:103625, 2025.
6. Szymon Bobek, Michał Kuk, Jakub Brzegowski, Edyta Brzychczy, and Grzegorz J. Nalepa. KnAC: an approach for enhancing cluster analysis with background knowledge and explanations. *Applied Intelligence*, 53:15537–15560, November 2022.
7. Szymon Bobek and Grzegorz J. Nalepa. Introducing uncertainty into explainable ai methods. In Maciej Paszynski, Dieter Kranzlmüller, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M. A. Sloot, editors, *Computational Science – ICCS 2021*, pages 444–457, Cham, 2021. Springer International Publishing.
8. Szymon Bobek and Grzegorz J. Nalepa. Local universal explainer (LUX) – a rule-based explainer with factual, counterfactual and visual explanations, 2023.
9. Szymon Bobek and Grzegorz J. Nalepa. Local universal rule-based explainer (LUX). *SoftwareX*, 30:102102, 2025.
10. Tapabrata Chakraborti, Christopher R. S. Banerji, Ariane Marandon, Vicky Hellon, Robin Mitra, Brieuc Lehmann, Leandra Bräuninger, Sarah McGough, Cagatay Turkay, Alejandro F. Frangi, Ginestra Bianconi, Weizi Li, Owen Rackham, Deepak Parashar, Chris Harbron, and Ben MacArthur. Personalized uncertainty quantification in artificial intelligence. *Nature Machine Intelligence*, 7:522–530, 2025.
11. Teodor Chiaburu, Felix Bießmann, and Frank Haußer. Uncertainty propagation in XAI: A comparison of analytical and empirical estimators. pages 390–411, 2026.
12. Dennis Collaris and Jarke J. van Wijk. ExplainExplore: Visual Exploration of Machine Learning Explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 26–35, June 2020. ISSN: 2165-8773.
13. Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking.
14. Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.
15. Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022. Preprint.
16. Anastasia-M. Leventi-Peetz and Kai Weber. Rashomon effect and consistency in explainable artificial intelligence (xai). In Kohei Arai, editor, *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1*, pages 796–808, Cham, 2023. Springer International Publishing.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29

17. Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, Jan 2020.
18. Dawid Macha, Michał Kozielski, Łukasz Wróbel, and Marek Sikora. Rulexai—a package for rule-based explanations of machine learning model. *SoftwareX*, 20:101209, 2022.
19. Maciej Mozolewski, Szymon Bobek, and Grzegorz J. Nalepa. Explaining time series classifiers with phar: Rule extraction and fusion from post-hoc attributions. *arXiv preprint*, 2025. Submitted.
20. Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the rashomon effect in explainable machine learning. In Danai Koutra, Claudia Plant, Manuel Gomez Rodriguez, Elena Baralis, and Francesco Bonchi, editors, *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 462–478, Cham, 2023. Springer Nature Switzerland.
21. Peyman Rasouli and Ingrid Chieh Yu. Explan: Explaining black-box classifiers using adaptive neighborhood generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.
22. Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.
23. Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, pages 1527–1535, New Orleans, Louisiana, USA, February 2018. AAAI Press.
24. Avi Schwarzschild, Max Cembalest, Karthik Rao, Keegan Hines, and John Dickerson. Reckoning with the disagreement problem: Explanation consensus as a training objective. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 662–678, New York, NY, USA, 2023. Association for Computing Machinery.
25. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, October 2017. ISSN: 2380-7504.
26. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_29