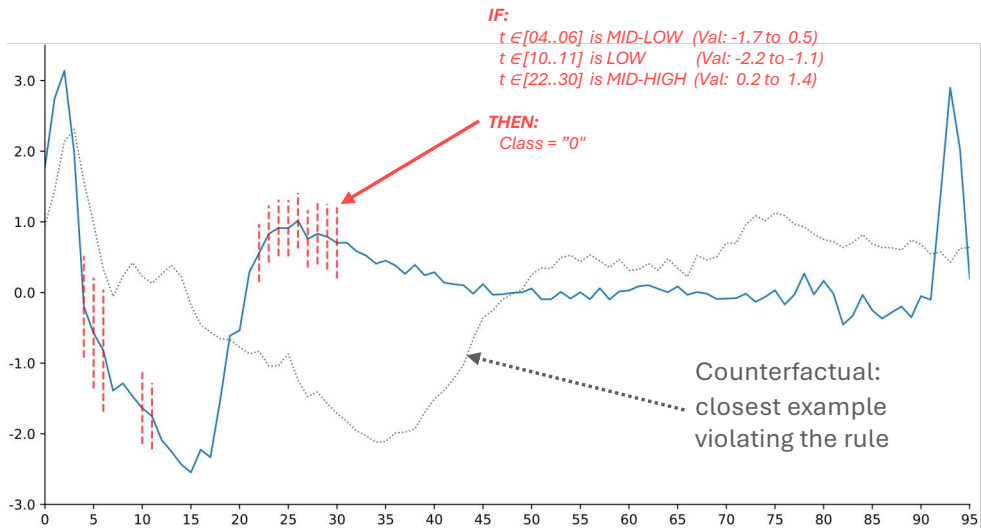
JAGIELLONIAN UNIVERSITY
IN KRAKÓW

peer

GEIST Research Group
We are GEIST. We dream big and work hard.

Beyond Heatmaps: Visualizing Decisions in Time Series with Interval Rules and Verifiable Counterfactuals

Maciej Mozolewski
Department of Human-Centered Artificial Intelligence
Institute of Applied Computer Science,
Faculty of Physics, Astronomy and Applied Computer Science,
Jagiellonian Univeristy



The Gap:
Local post-hoc XAI vs. Time Series Reality

The Drawbacks of Explanations

Abstract Heatmaps: Show *where* the model looks, but fail to specify *what* signal values are needed.

The Disagreement Problem: Strong temporal autocorrelation (SHAP, LIME) or sparsity optimisation (Anchor) causes explainers to conflict.

Domain Disconnect: Domain experts require structural shape constraints, not abstract importance scores.

What Experts Actually Need

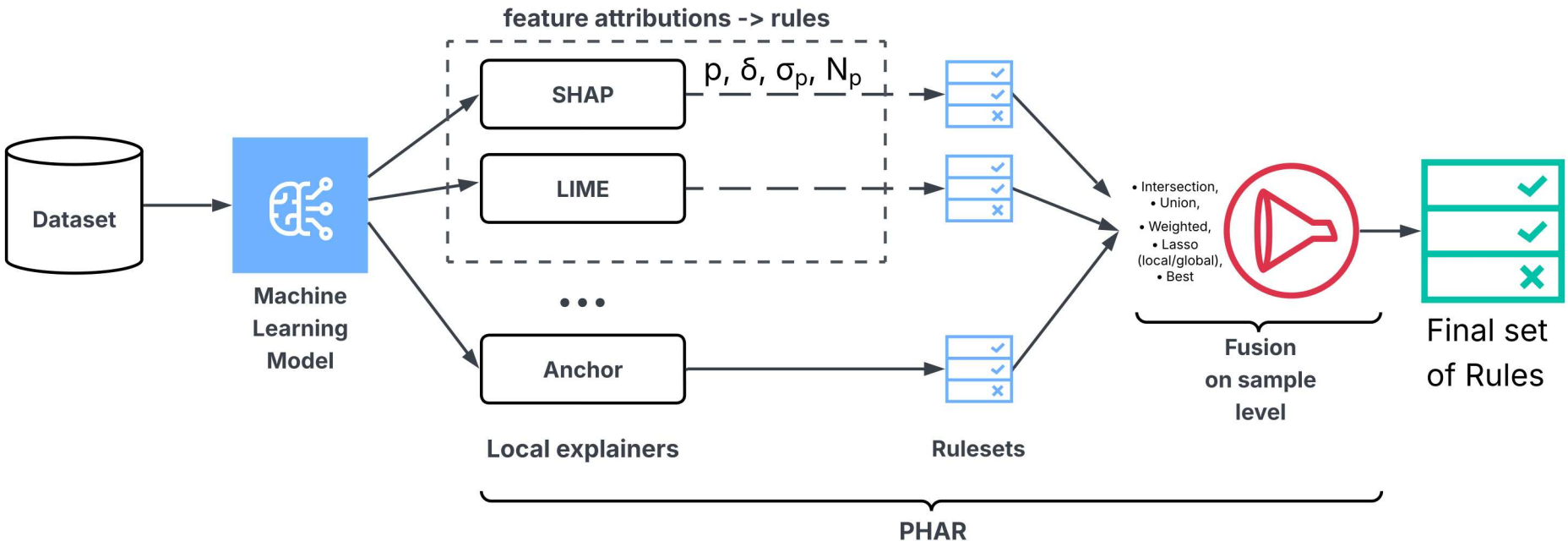
Symbolic Translation: Semi-factual intervals defining explicit, stable bounds *(Solution:* Post-hoc Attribution Rules*)*.

Rule Consensus: Formal ambiguity resolution mechanisms to filter noise and resolve contradictions *(Solution: Rule fusion)*.

"What-If" scenarios: Support for expert-in-the-loop curation *(Solution: Verifiable Counterfactuals)*.

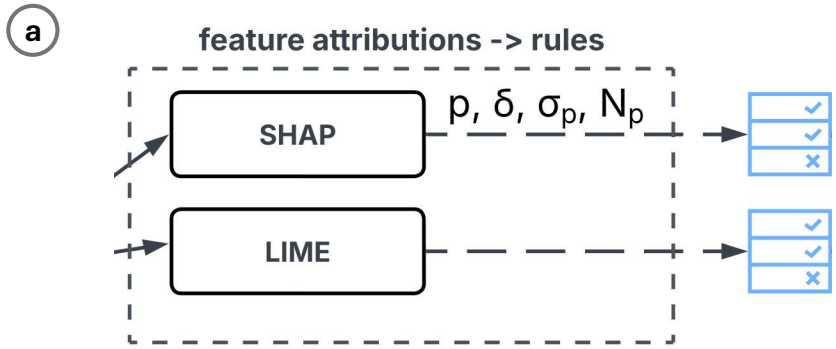
PHAR: Post-hoc Attribution Rules for Time Series

1



M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

PHAR: Translating Attributions into Semi-Factual Intervals



Algorithm: PHAR Extraction

Inputs:
- x : Input instance
- $y^$: Original model prediction for x
- e : Numeric feature attributions (e.g., local SHAP/LIME values)
- std_f : standard deviations of features in train set

Hyperparameters (tuned via Bayesian optimization):
- p : Percentile threshold
- δ : Boolean flag (Global vs. Local thresholding)
- σ_p : Perturbation scale
- N_p : Number of perturbation samples

1. // Define Explanation Multisets (across training data)
   $E = \{ |e_{n,f}| \text{ for all instances } n, \text{ features } f \}$
   $E_f = \{ |e_{n,f}| \text{ for all instances } n, \text{ specific feature } f \}$

2. // Feature Selection
   IF δ is True:
   $T = \text{Percentile}(E, p)$
   Select important features F^* where $|e_f| \geq T$
   ELSE:
   For each f: $T_f = \text{Percentile}(E_f, p)$
   Select important features F_* where $|e_f| \geq T_f$

3. // Joint Perturbation & Sampling
   FOR 1 to N_p:
   For each $f \in F_*$:
   Sample $x'_f \sim \text{Uniform}(x_f - \sigma_p * std_f, x_f + \sigma_p * std_f)$
   Evaluate model prediction for perturbed sample x'

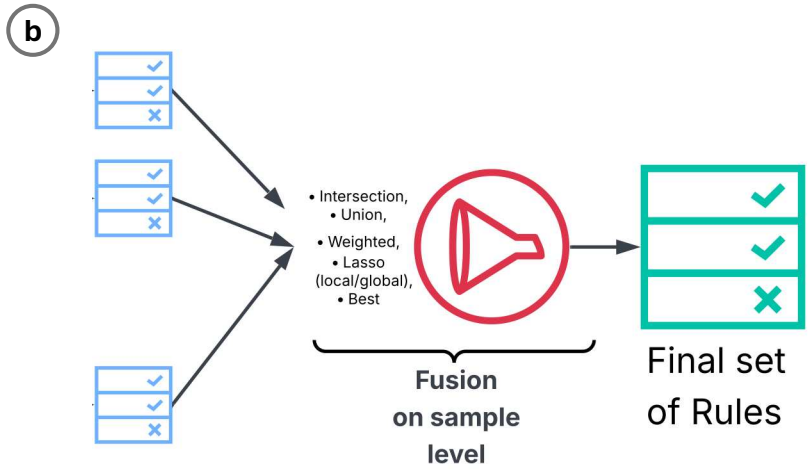
4. // Interval Derivation
   Find stable bounds $(l_f, u_f]$ for all $f \in F_*$ preserving prediction $y^$

5. // Rule Quality Evaluation
   Evaluate bounds on the rule extraction set:
   COV = % of instances satisfying the rule bounds
   CONF = % of covered instances matching prediction $y^$

6. RETURN Rule R_n:
   IF: $\bigwedge_{\{f \in F_*\}} : x_f \in (l_f, u_f]$
   THEN: Class = $y^$, (CONF, COV)

M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

PHAR: Resolving the Rashomon Effect via Formal Rule Fusion



Algorithm: Rule Fusion Across Explainers

Inputs:
- x_i : Input instance
- R_M : Set of candidate rules for x_i from explainers M (e.g., {SHAP, LIME, Anchor})
- S : Fusion Strategy (Intersection, Union, Weighted, Lasso/ Lasso Global, Best)

1. // Feature Selection & Interval Aggregation
   MATCH Strategy S:

   CASE **'Best'**:
   RETURN single rule $R \in R_M$ maximizing metric (e.g., CONF)

   CASE **'Intersection'**:
   F_{fused} = Features present in ALL candidate rules
   $Bounds_f$ = [$\max(l_f)$, $\min(u_f)$] for each $f \in F_{fused}$

   CASE **'Union'**:
   F_{fused} = Features present in ANY candidate rule
   $Bounds_f$ = [$\min(l_f)$, $\max(u_f)$] for each $f \in F_{fused}$

   CASE **'Weighted'**:
   F_{fused} = Features where weighted presence > threshold τ
   $Bounds_f$ = [$\min(l_f)$, $\max(u_f)$] for each $f \in F_{fused}$

   CASE **'Lasso'** (Local / Global):
   // Local: fits only x_i. Global: fits ALL instances.
   Encode rule conditions as binary matrix Z
   Solve sparse regression: min $||y - Z\beta||^2 + \lambda|\beta|$
   F_{fused} = Retain cond. where $\beta \approx 0$ (intervals merged by union)

2. // Construct Fused Rule (if S is not 'Best')
   R_{fused}: $\bigwedge_{\{f \in F_{fused}\}} : x_f \in Bounds_f\ (l, u]$

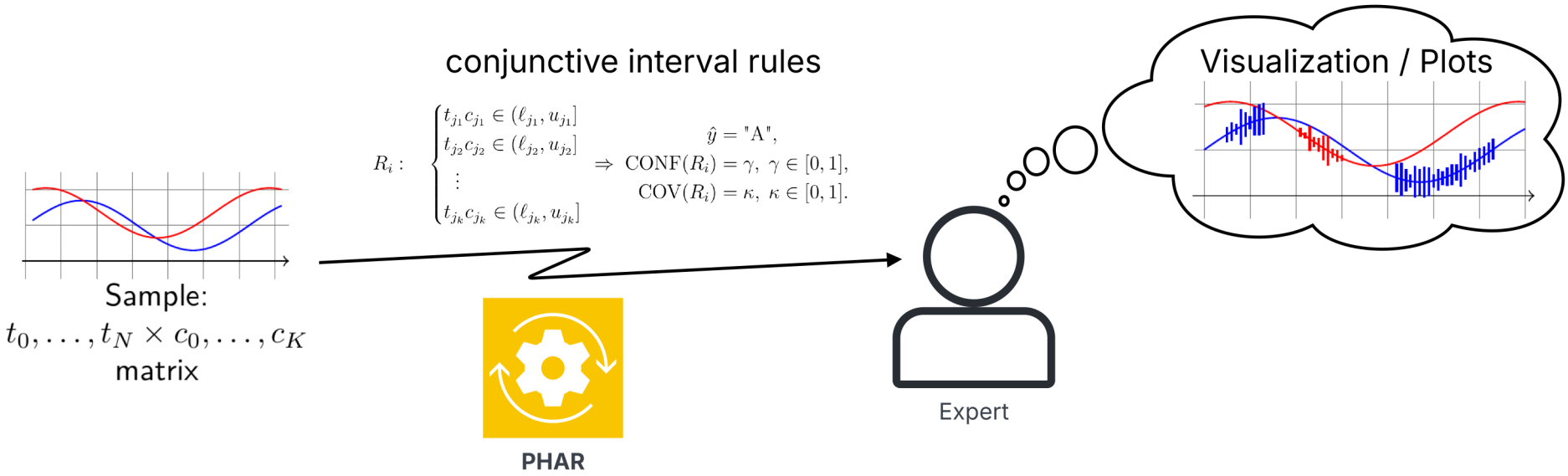
3. // Rule Quality Re-Evaluation
   Evaluate R_{fused} on the rule extraction set :
   COV = % of instances satisfying R_{fused} bounds
   CONF = % of covered instances matching original prediction y^

4. RETURN Fused Rule (with CONF, COV)

M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

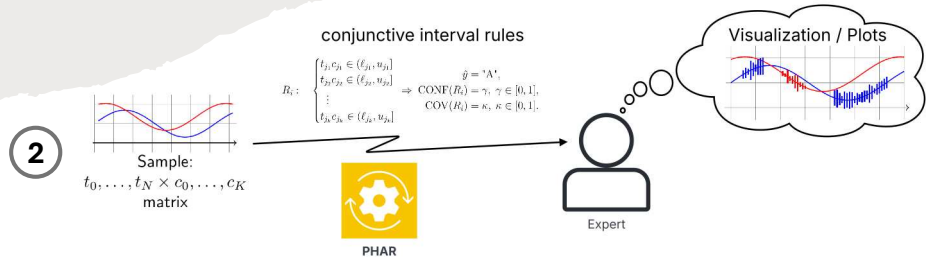
PHAR: Local Interval Rules and Visualisations

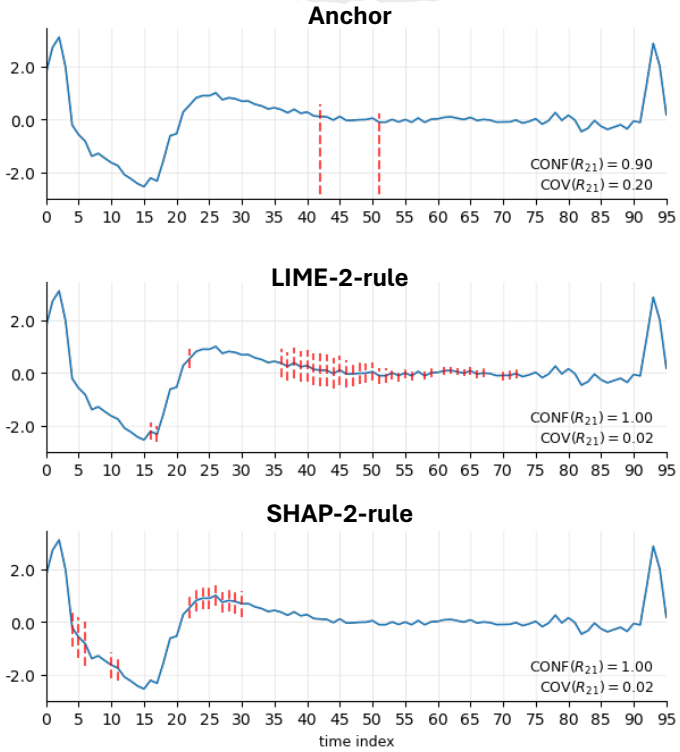
2



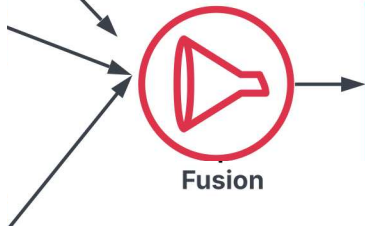
M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

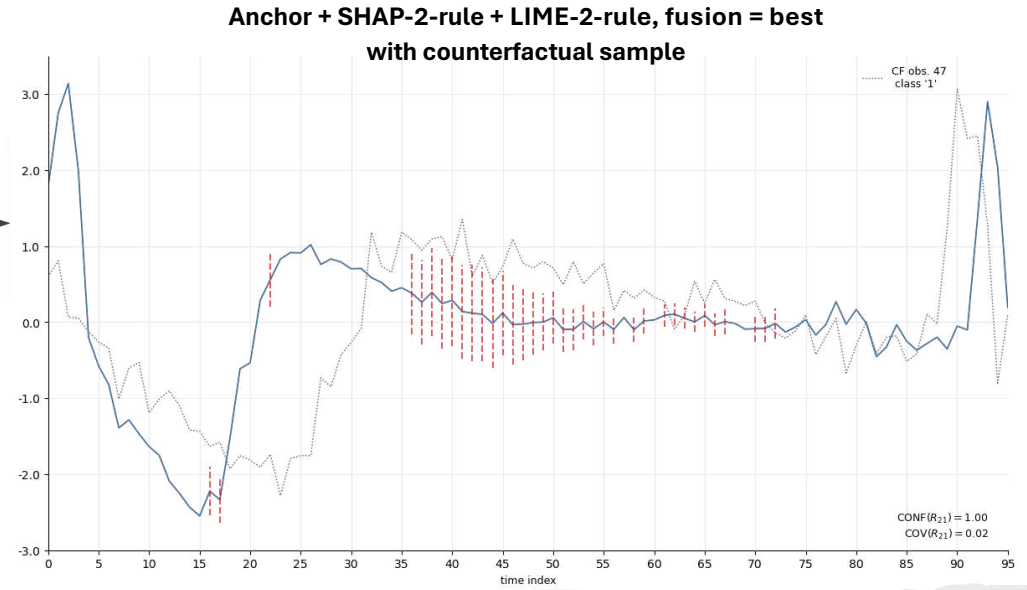
Verification with Counterfactuals





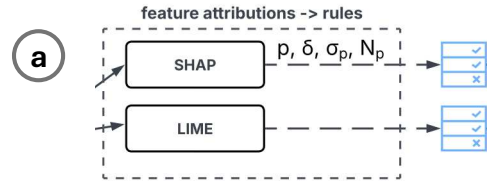
Dataset: ECG200 (univariate), sample 21





M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

Quantitative Evaluation: The Sparsity vs. Shape Trade-off



Experimental Setup: Evaluated on 43 diverse TS datasets (UCR/UEA) using Deep ConvLSTM1D classifiers.

Optimisation of PHAR extraction

Hyperparam.	Range/Values	Description
p	50 to 99 (step=1)	Percentile for determining feature importance thresholds.
δ	{True, False}	Global vs. per-feature threshold application.
σ_p	0.01 to 1.0 (continuous)	Scale of perturbation for interval estimation.
N_p	1,000 to 10,000 (step=1,000)	Number of perturbed samples for interval estimation and confidence evaluation.

p - threshold percentile, δ - global importance indicator, σ_p - perturbation scale, N_p - perturbation samples count

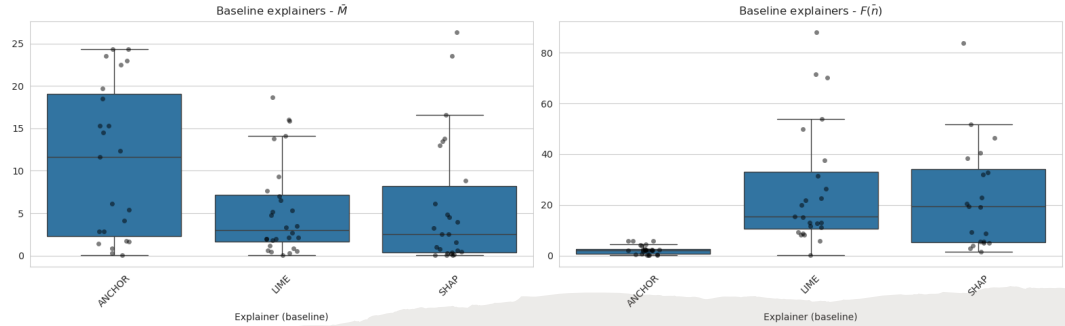
Base Score: Maximizes the trade-off between rule precision (Confidence) and generality (Coverage).
Penalty Constraints: Soft constraints explicitly reject low-quality or overly complex logic.

$$M(R_n) = \underbrace{\left(\gamma(R_n) \cdot \kappa(R_n)\right)}_{\text{Base Score}} \cdot \mathscr{P}_{conf} \cdot \mathscr{P}_{cov} \cdot \mathscr{P}_{len} \qquad \bar{M}(\theta) = \frac{1}{N}\sum_{n=1}^{N} M(R_n(\theta))$$

$$\mathscr{P}_{conf} = \min\left(1, \frac{\gamma(R_n)}{\tau_{conf}}\right), \quad \mathscr{P}_{cov} = \min\left(1, \frac{\kappa(R_n)}{\tau_{cov}}\right), \quad \mathscr{P}_{len} = \min\left(1, \frac{\tau_{len}}{F(n)}\right)$$

High Fidelity: SHAP-derived interval rules achieved a median CONF of **94%**, outperforming both native Anchor (**87%**) and LIME-derived (**87%**).

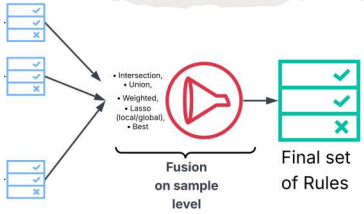
Sparsity:
- Anchor yields very sparse rules (median **2.0** features) and relies on one-sided bounds.
- SHAP (median **4.5**) and LIME (median **5.4**) use more features to create precise interval bounds (l_f, u_f], better specifying temporal segments.



M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

Quantitative Evaluation: The Sparsity vs. Shape Trade-off

b

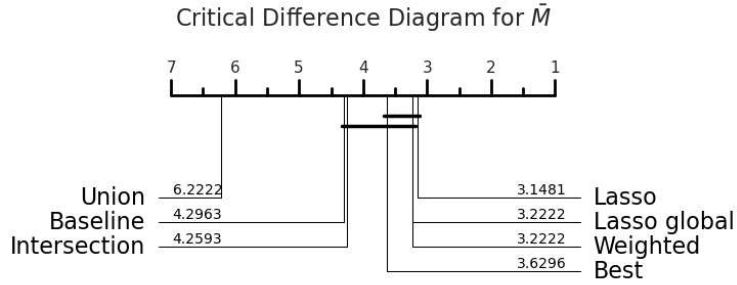


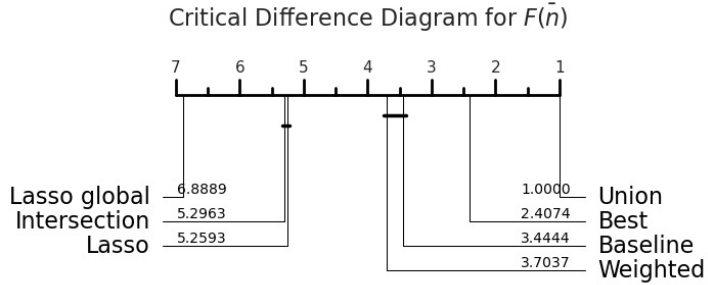
The Rashomon Effect: Different explainers yield conflicting valid boundaries. Naively aggregating them (*Union*) captures shape but causes severe rule bloating (median **42.6** features), destroying interpretability.

Fusion strategies: *Lasso*, *Weighted*, *Best* consistently top the non-parametric evaluation (Friedman & Nemenyi post-hoc tests).

The "Best" Strategy: Simply selecting the top-performing rule per instance achieves the highest median Confidence (**97%**), with low sparsity (median **26.0** features).

The Lasso (Local): maintains high fidelity (median Confidence: **94%**) and Coverage (**12%**) while keeping cognitive load manageable at a median of **6.0** features per rule.





M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

The Gap: From Local Explanations to Deployable Surrogates

The Drawbacks of Local Extractors

Fragmented Logic: Local extractors provide isolated explanations, lacking a cohesive global overview.

Inference-Time Conflicts: Post-hoc rules are not true surrogates; they frequently overlap and contradict each other on new data.

Explainer Lock-in: Relying on a single induction method limits flexibility and yields suboptimal boundaries.

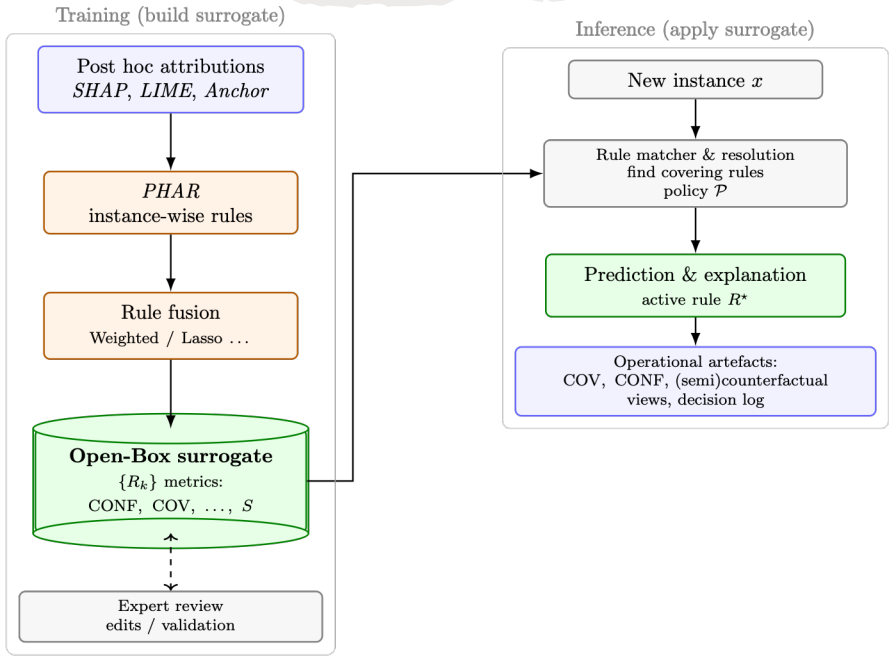
The Open-Box Solution

Global Glass-Box: Aggregating instance-level rules into a unified, auditable surrogate *(Solution:* Open-Box*)*.

Explicit Resolution: Inference-time mechanisms to systematically determine the final prediction *(Solution:* Resolution Modes*)*.

Agnostic Integration: Seamlessly importing logic from diverse SOTA sources like PHAR, LUX, or Anchor *(Solution:* Modular Architecture*)*.

Open-Box: Explainer-Agnostic Global Surrogate



Open-Box

- Aggregates local rules (PHAR, LUX, Anchor, ...) into **Global Surrogate**.
- Resolves overlapping rules on **unseen data** via explicit inference policies (e.g., *highest confidence*).

Inference-Time Resolution Modes

- Resolves overlapping rules on **unseen data** via explicit resolution policies (Score-based Selectors; Geometric Selectors; Voting).

Resolution Mode	Selection Criterion	Practical Impact
`highest confidence`	Maximizes empirical precision, $c(r)$	Prioritizes rules with the highest historical reliability, minimizing false positives.
`confidence` × `coverage`	Maximizes the product $c(r) \cdot v(r)$	Balances reliability with broad applicability, avoiding overfitting to overly narrow, idiosyncratic rules.
`narrow`	Minimizes average interval width	Favors highly specific rules that tightly bound the local feature space.
`central`	Minimizes normalized distance to rule bounding box center	Selects rules where the new instance lies deepest within the established safe decision boundary.
`last` / `first`	Deterministic tie-breaking based on extraction sequence	Serves as a fast, computationally cheap fallback relying on the order of rule generation.
`voting` (variants)	Aggregates predictions uniformly or weighted by metric	Leverages ensemble effects to reduce variance, providing robust decisions when local rules conflict.

M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Open-Box: Extracting Interpretable Rules from any classifiers*" (2026), Submitted to *IEEE Access*.

Evaluation: OpenML Benchmarks & Industrial Case Study

OpenML

- **Experimental Setup**: 37 curated datasets. Strict 60/30/10 data split (Training / Rule Extraction / Hold-out evaluation).
- **Model Baselines**: Compared canonical white-boxes (Decision Tree, Logistic Regression, k-NN) against black-boxes (LightGBM, MLP, Random Forest, FlowVectorClassifier).
- **Results**: The surrogate achieved near-saturated coverage 99%) and high covered accuracy (90%) on the unseen hold-out sets.
- Geometric Selectors should be used for models of high Accuracy.

The "White-Box" Illusion

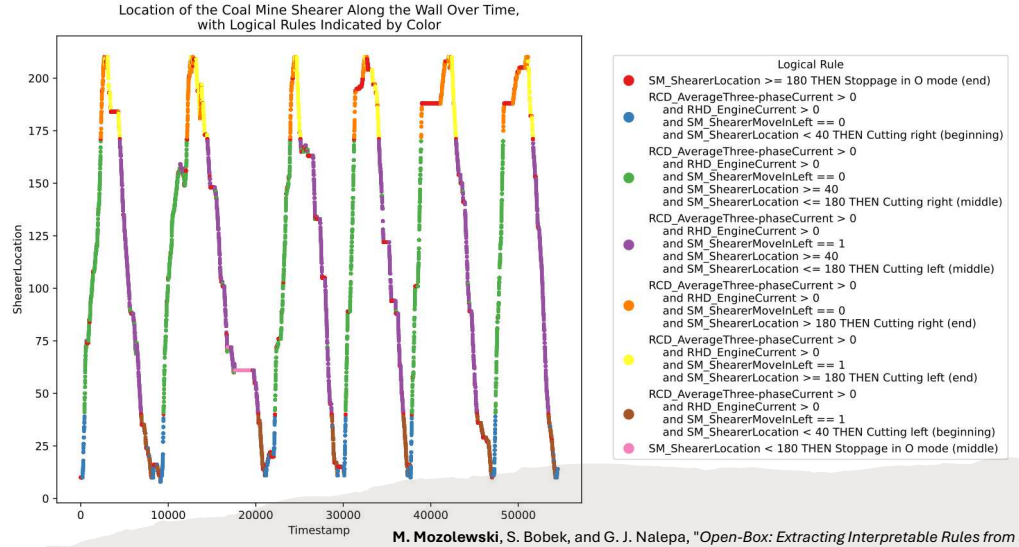
```
kNN explanation
• k=5, weights=uniform, metric=minkowski
• predicted label: P
• top-5 neighbors (smaller distance → more influence):
 rank  train_index  distance label  weight
    1         2690  0.001021     P     1.0
    2         1775  0.001085     P     1.0
    3         1121  0.001095     P     1.0
    4         1310  0.001213     P     1.0
    5         1041  0.001223     P     1.0

Nearest neighbor (rank 1) -- feature vector:
RollRate        0.0047
PitchRate      -0.0010
currPitch       0.0060
currRoll       -0.0070
diffRollRate    0.0001
```

```
{
  "index": 95,
  "rule_kind": "intervals",
  "rule": "RollRate >0.003386350184527866 &
     <=0.004613466197360241",
  "confidence": 0.9105058365758756,
  "coverage": 0.1201496026180458,
  "exp_count": 1,
  "matched_rule_count": 1,
  "openbox_source_model": "kNN",
  "shap2rulesPerturbation": "GlobalTrue",
  "shap2rulesPercentile": "90",
  "resolution_mode": "last",
  "classes": {
          "ground_true": "P",
          "openbox_rule": "P",
          "whitebox_model": "P"/"P"/"P" }
}
```

Coal Mine Case Study (9 classes)

- **Expert Rules:** 61.5% accuracy, **Black-Box** (LightGBM): 88.4% accuracy.
- Local Extractors (Accuracy):
  - **LUX**: 87.8% (with linear formulas e.g., $A < B*c_1 + c_2$)
  - **PHAR**: 83.3% • **Anchor**: 82.7% • **EXPLAN**: 51.1% • **LORE**: 48.5%
- **Open-Box Surrogate** (LIME-based)**:** 88.0%.
- *Result:* Extracted tacit domain knowledge experts missed.



M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Open-Box: Extracting Interpretable Rules from any classifiers*" (2026), Submitted to *IEEE Access*.

The Gap:
Standard Counterfactuals vs. Clinical Reality

The Drawbacks of Standard CFs in Healthcare

Physiological Unreality: Standard counterfactuals apply dense, global changes, creating unrealistic and clinically implausible signals.

Temporal Misalignment: Explanations often ignore the natural rhythms of biological signals, leading to mismatched phases (e.g., broken heartbeats).

Clinical Disconnect: Abstract, slow-to-generate perturbations fail to support rapid, expert-driven medical diagnosis.

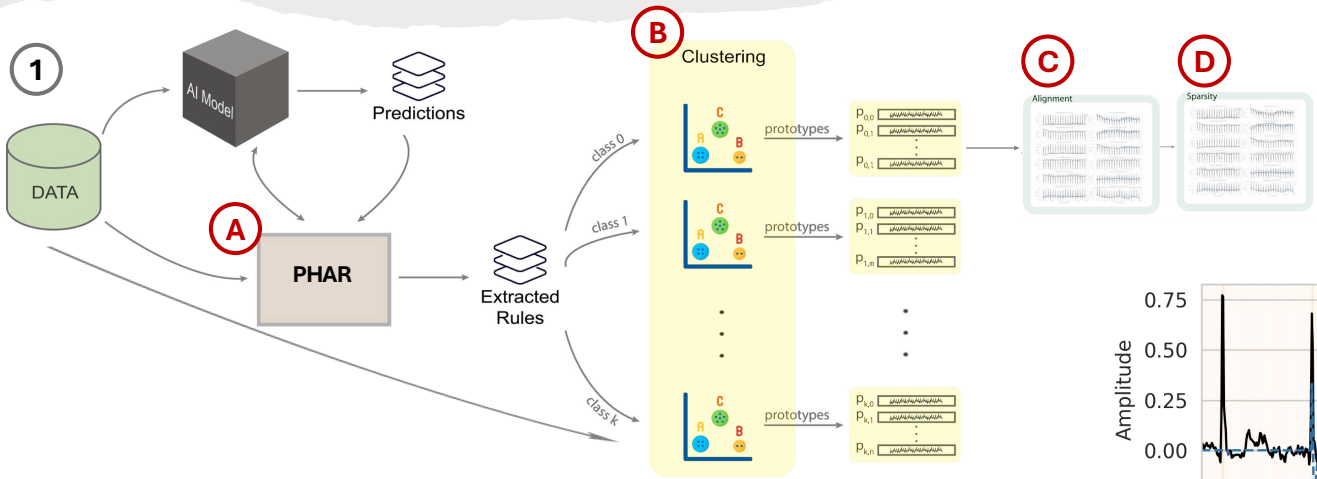
The Sparse ECG Solution

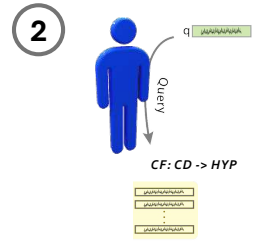
Targeted Sparsity: Modifying only critical segments guided by SHAP-derived interval rules to preserve natural morphology *(Solution: Sparse Counterfactuals)*.

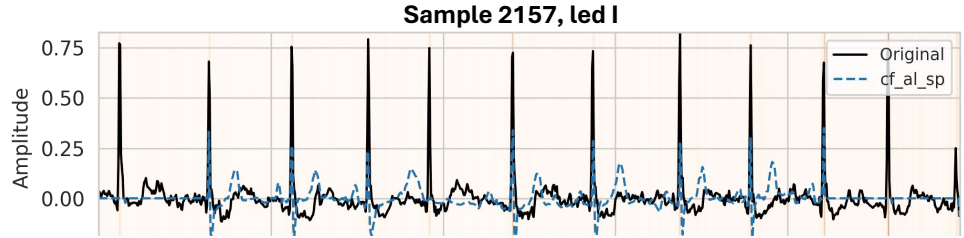
Physiological Alignment: Using DTW clustering to extract real prototypes and strictly align them to the patient's R-peaks *(Solution: Prototype Alignment)*.

Actionable & Fast: Near real-time (< 1s) generation of physiologically valid scenarios designed for interactive clinical platforms *(Solution:* Expert-Informed Design*)*.

Sparse ECG Counterfactuals







Step A: PHAR: Interval rules to identify the most critical ECG segments.

Step B: DTW & Prototypes: Extracts real-patient medoids as class prototypes via Dynamic Time Warping (DTW) distances and Multi-Dimensional Scaling (MDS) clustering.

For a new query:

Step C: R-peak Alignment: Temporally warps the prototype to match the query ECG's exact cardiac rhythm.

Step D: Sparsity Optimization: Iteratively swaps minimal critical segments (highlighted by SHAP) towards the prototype to successfully flip the prediction.

- **PTB-XL:** 12-lead ECG, 5 diagnostic classes: NORM, MI, STTC, CD, HYP.
- **Validity: 81.3%** overall success in flipping predictions (**98.9%** for MI).
- **Stability:** Temporal stability *(robustness to small time shifts)* improved by **43%** vs. unaligned prototypes (0.5304 vs 0.7593).
- **Speed:** Pre-computed prototypes enable real-time generation (**< 1 second**).
- **Expert Validation:** Cardiology experts provided positive initial feedback on clarity, readability, and reduced cognitive load.

The Gap:
Abstract Attributions vs. Anatomical Reality

The Drawbacks of Current ECG XAI

Abstract Localization: Standard 12-lead heatmaps highlight abstract waveform fluctuations, forcing clinicians to mentally map them to physical heart structures.

Spatial Model Degradation: Training models directly on 3D spatial data (CineECG) degrades predictive accuracy and yields incoherent, noisy explanations.

Attribution Instability: Gradient-based temporal heatmaps frequently collapse (e.g., at the isoelectric line) or fade across low-amplitude features.

The Cross-Modal Solution

Anatomical Projection: Projecting 1D feature importance directly onto the 3D CineECG ventricular model for intuitive spatial reasoning *(Solution: Cross-Modal Mapping)*.

Best of Both Worlds: Using a robust 12-lead model as the primary feature extractor, inheriting its high accuracy while gaining 3D explanatory clarity *(Solution: 12-Lead to 3D Transfer)*.

Spatial Regularization: The 3D anatomical integration acts as a natural filter, smoothing temporal noise and presenting a cohesive pathological region *(Solution: Anatomical Filtering)*.

Validating CineECG via Cross-Modal XAI

The Goal: Validating feature attributions against clinical ground truth (expert annotations).

[Expert Annotation Platform]

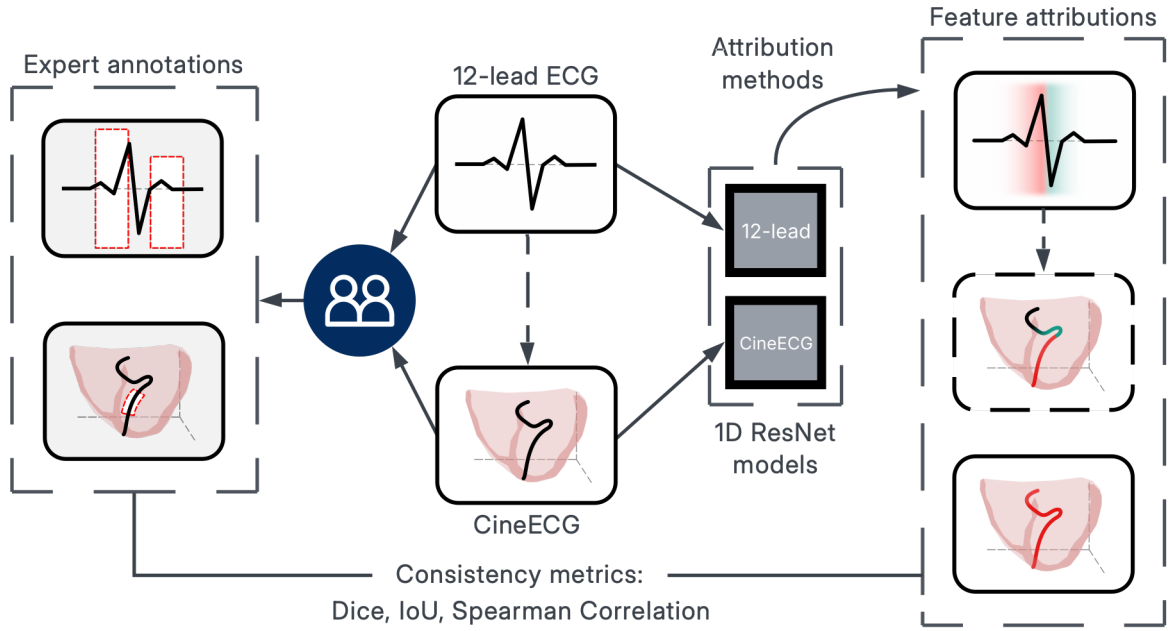
Models: ResNet on 8,000 temporally aligned 12-lead and CineECG records (normal/healthy vs abnormal).

The Performance Gap: Standard 12-lead models (**83.1% Acc**) beat direct 3D CineECG training (**79.1% Acc**).

The Solution: Mapping 1D temporal attributions (Integrated Gradients) directly onto the 3D anatomy.

XAI validation: cardiologists blindly annotated 20 independent test cases to define binary ground-truth masks.

Spatial Regularization: 3D projection filters temporal noise, increasing alignment with experts (Dice: 0.47 → 0.56)



K. Dobiczek & **M. Mozolewski** *(equal contribution)*, S. Bobek, M. Szafarczyk, P. van Dam, and G. J. Nalepa, *"Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution"* (2026), Submitted to ICCS.

Validating CineECG via Cross-Modal XAI

The Goal: Validating feature attributions against clinical ground truth (expert annotations).

[Expert Annotation Platform]

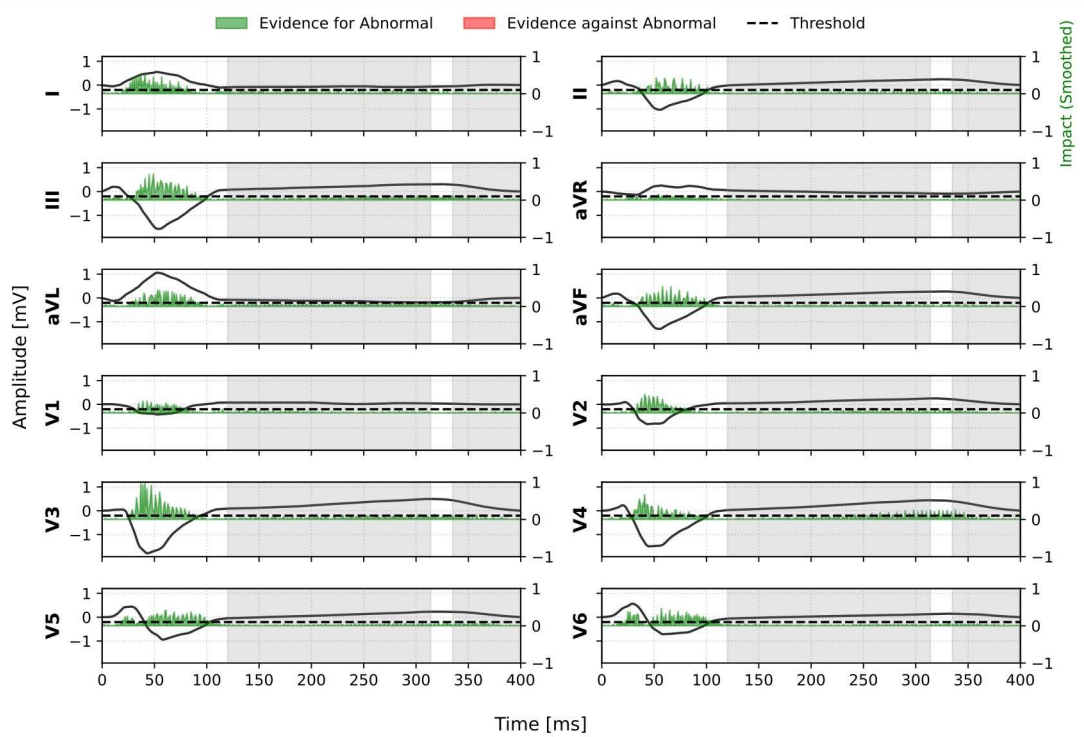
Models: ResNet on 8,000 temporally aligned 12-lead and CineECG records (normal/healthy vs abnormal).

The Performance Gap: Standard 12-lead models (**83.1% Acc**) beat direct 3D CineECG training (**79.1% Acc**).

The Solution: Mapping 1D temporal attributions (Integrated Gradients) directly onto the 3D anatomy.

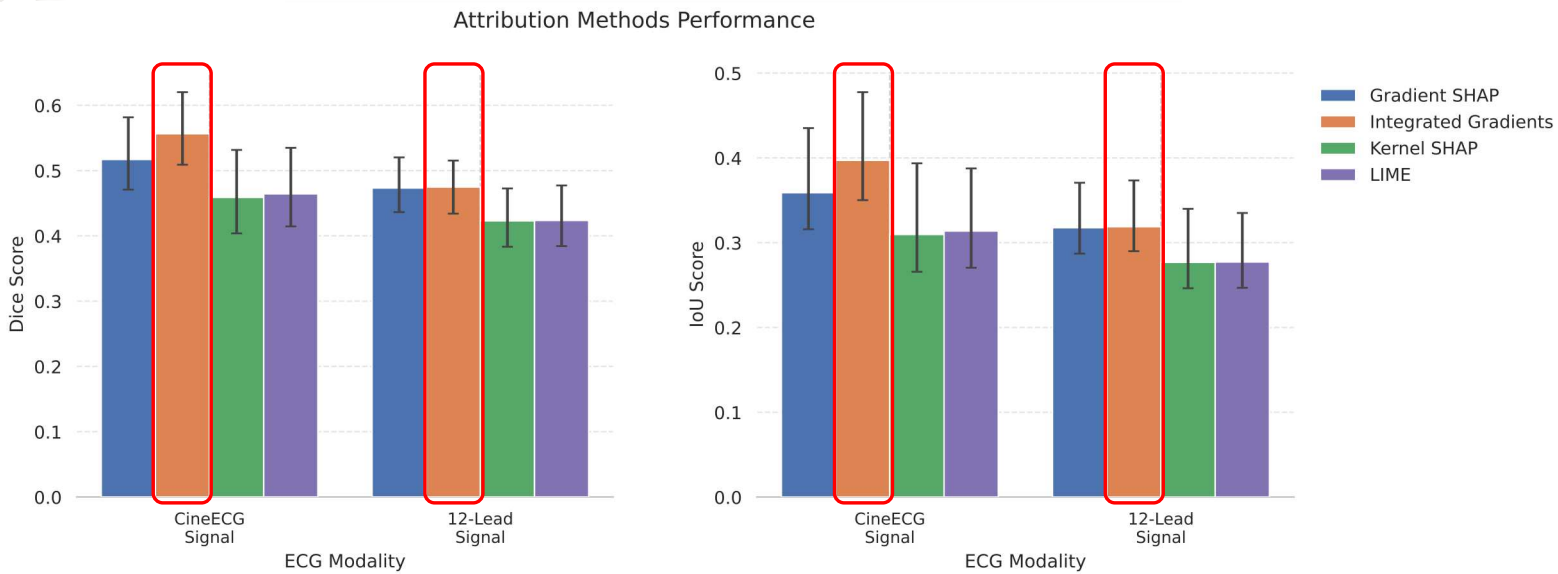
XAI validation: cardiologists blindly annotated 20 independent test cases to define binary ground-truth masks.

Spatial Regularization: 3D projection filters temporal noise, increasing alignment with experts (Dice: 0.47 → 0.56)



K. Dobiczek & **M. Mozolewski** *(equal contribution)*, S. Bobek, M. Szafarczyk, P. van Dam, and G. J. Nalepa, *"Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution"* (2026), Submitted to ICCS.

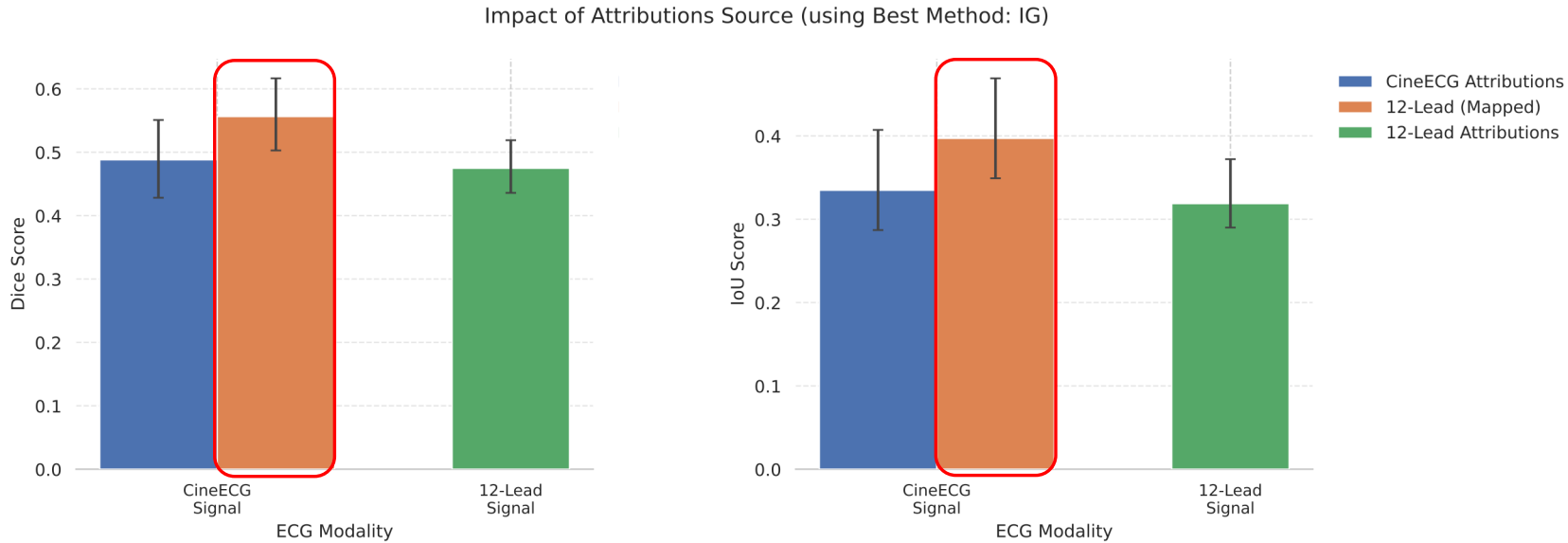
Best XAI method: Integrated Gradients



- *Integrated Gradients (IG)* yields the best spatial overlap when projected to 3D space (Dice: 0.47 vs 0.56; IoU: 0.32 vs 0.40).
- Perturbation methods (LIME, Kernel SHAP) obtained lower scores regardless of the modality

K. Dobiczek & **M. Mozolewski** *(equal contribution)*, S. Bobek, M. Szafarczyk, P. van Dam, and G. J. Nalepa, *"Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution"* (2026), Submitted to ICCS.

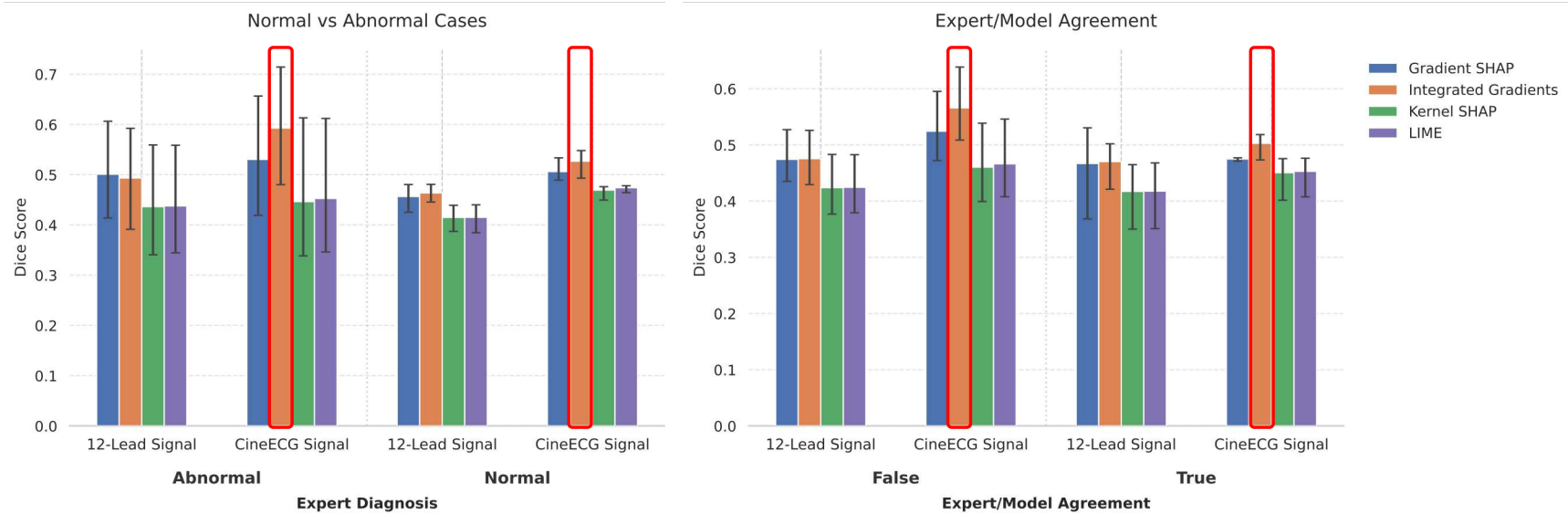
Best XAI modality: mapping 12-led ECG to Cine3D



- Mapping 12-led attributions to 3D space directly improves alignment with cardiologists (**Dice: 0.56 vs 0.47**; **IoU: 0.40 vs 0.32**)
- The anatomical 3D model successfully acts as a natural noise filter for temporal attributions.

K. Dobiczek & **M. Mozolewski** *(equal contribution)*, S. Bobek, M. Szafarczyk, P. van Dam, and G. J. Nalepa, *"Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution"* (2026), Submitted to ICCS.

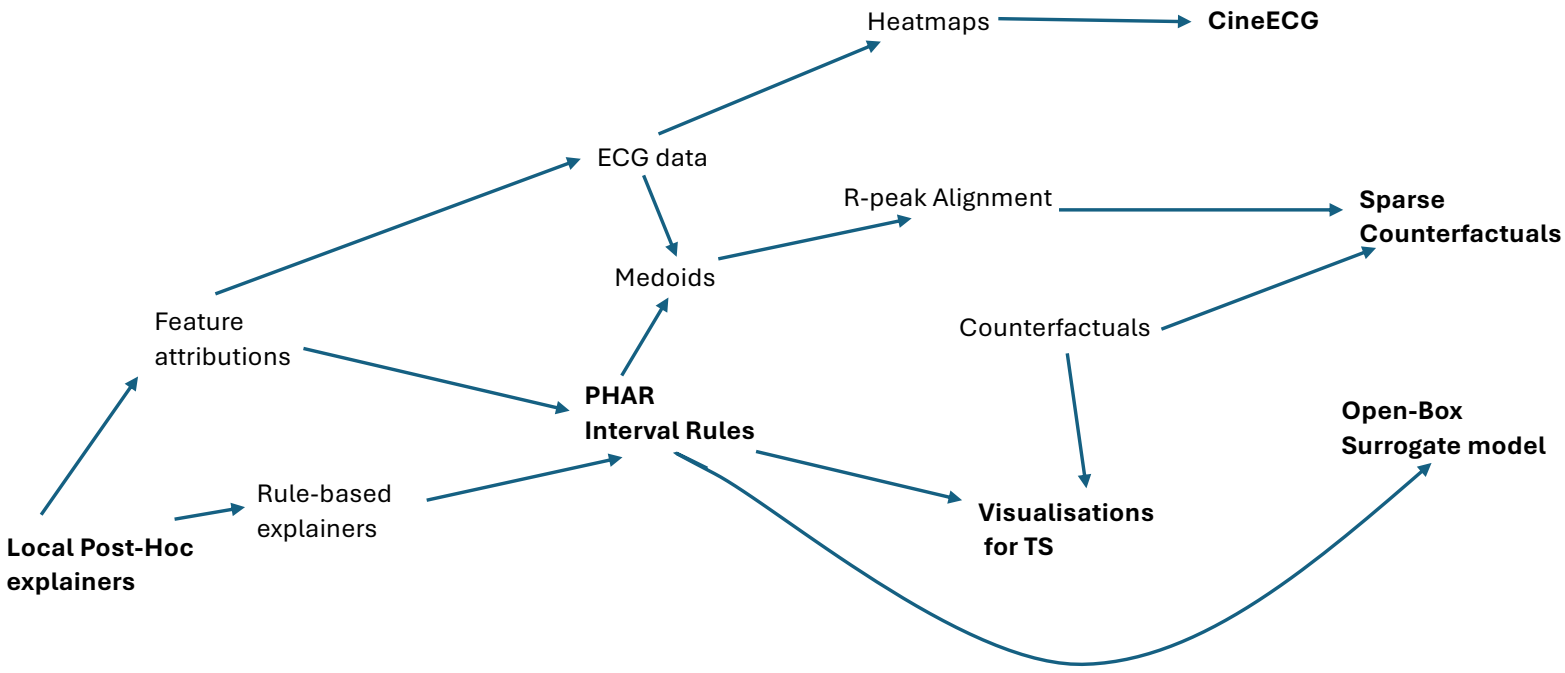
Best agreement for Abnormal cases & where model does not agree



- **Pathology:** Higher alignment on *Abnormal* cases (**Dice: 0.59**) than *Normal* cases (**0.53**). Focal anomalies are easier to localize than healthy signals.
- **Disagreement:** Mapped 3D explanations align *better* with clinical experts when the model classifies *incorrectly* (**Dice: 0.57 vs 0.50**).
- **Debugging:** Cross-modal mapping correctly isolates pathological morphology, even when the model's final classification head fails.

K. Dobiczek & **M. Mozolewski** *(equal contribution)*, S. Bobek, M. Szafarczyk, P. van Dam, and G. J. Nalepa, *"Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution"* (2026), Submitted to ICCS.

Summary



M. Mozolewski & B. Bayrak *(equal contribution)*, K. Bach, and G. J. Nalepa, *"From Prototypes to Sparse ECG Explanations: SHAP-Driven Counterfactuals for Multivariate Time-Series Multi-class Classification"* (2026), Information Systems Frontiers. DOI: 10.1007/s10796-026-10711-9.

M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Open-Box: Extracting Interpretable Rules from any classifiers*" (2026), Submitted to *IEEE Access*.

K. Dobiczek & **M. Mozolewski** *(equal contribution)*, S. Bobek, M. Szafarczyk, P. van Dam, and G. J. Nalepa, *"Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution"* (2026), Submitted to ICCS.

Summary

Key Takeaways

PHAR: Translated abstract numeric heatmaps into continuous, semi-factual interval rules and visualises them with countervatulas.

Open-Box: Fused fragmented local rules into a deployable global model, delivering interpretable explanations and matching black-box fidelity in industrial applications.

Prototype-Driven ECG Counterfactuals: Leveraged prototypical parts and SHAP and R-peak alignmen to generate sparse, actionable explanations for multivariate ECG classification.

CineECG: Proved that 3D anatomical projection acts as a natural noise filter, boosting alignment with clinical experts.

Future Work

Interactive Curation: Implementing *literal-level importance* scoring to actively guide domain experts during manual rule refinement.

Clinical Validation: Positive preliminary feedback establishes a strong baseline, but large-scale clinical trials are needed.

High-Stakes Deployment: Scaling the unified framework for continuous clinical trials and dynamic industrial monitoring.

M. Mozolewski & B. Bayrak *(equal contribution)*, K. Bach, and G. J. Nalepa, *"From Prototypes to Sparse ECG Explanations: SHAP-Driven Counterfactuals for Multivariate Time-Series Multi-class Classification"* (2026), Information Systems Frontiers. DOI: 10.1007/s10796-026-10711-9.

M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Time Series Explainability with Post-hoc Attribution Rules, Semi-Factual Intervals and Verifiable Counterfactuals*" (2026).

M. Mozolewski, S. Bobek, and G. J. Nalepa, "*Open-Box: Extracting Interpretable Rules from any classifiers*" (2026), Submitted to *IEEE Access*.

K. Dobiczek & **M. Mozolewski** *(equal contribution)*, S. Bobek, M. Szafarczyk, P. van Dam, and G. J. Nalepa, *"Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution"* (2026), Submitted to ICCS.

JAGIELLONIAN UNIVERSITY IN KRAKÓW

peer

GEIST Research Group

Thank you for your attention!

This research is supported by **Horizon Europe project PEER** – *The Hyper-Expert Collaborative AI Assistant* (Grant Agreement No. 101120406).
More details peer-ai.eu.