4th World Conference on eXplainable AI · Fortaleza, Brazil · July 2026

A Structured Guide to Selecting Feature Attribution Techniques

Priscylla Silva · Victor H. Barella · Luis Gustavo Nonato

Federal Institute of Alagoas · ICMC, University of São Paulo

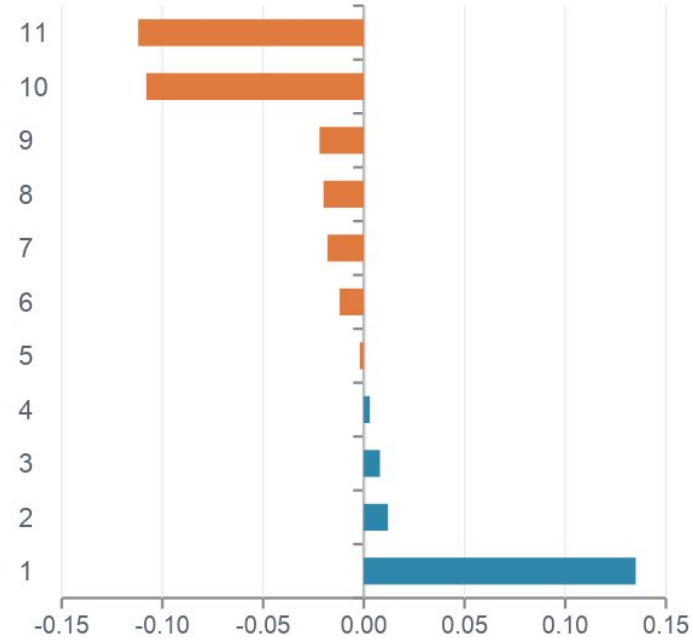
MOTIVATION

Too many methods, no way to choose

Feature attribution methods assign scores to input features — revealing why a model made a prediction. But the field has produced dozens of methods, and selecting one is genuinely hard:

Disagreement. Different methods give conflicting explanations for the same prediction.

No standard criteria. Practitioners often choose by popularity or intuition, not evidence.



LIME explanation — heart-disease model

Blue increases · orange decreases disease likelihood

Three intuitive approaches — all flawed

A · Pick one method

"Just use LIME, it's popular."

- No justification
- May misalign with model
- Better options ignored
- Faithfulness unknown

B · Ask an expert

Let a cardiologist judge.

- Subjective & biased
- Experts disagree
- Faithfulness unknown
- Cognitive overload

C · Use metrics

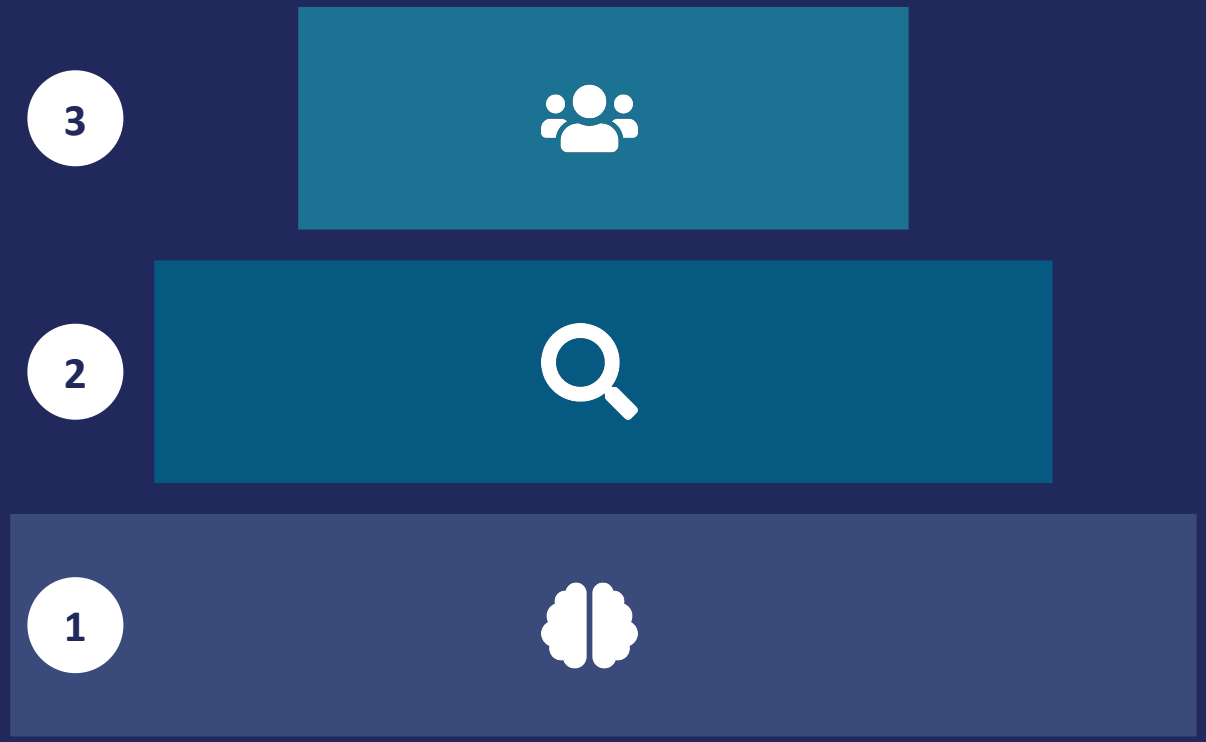
Rank by quantitative scores.

- Metrics disagree
- May miss clinical need
- Property trade-offs

Arbitrary, expert-only, and metric-only selection each break down. We need a structured method that combines all three.

OUR CONTRIBUTION

A pyramid framework for method selection



Human-Centric

Useful, trustworthy & understandable to end users

Explanation-Centric

Robust, stable & low-complexity explanations

Model-Centric

Explanations faithfully reflect the model

Bottom-up: each layer filters methods before the next, costlier evaluation

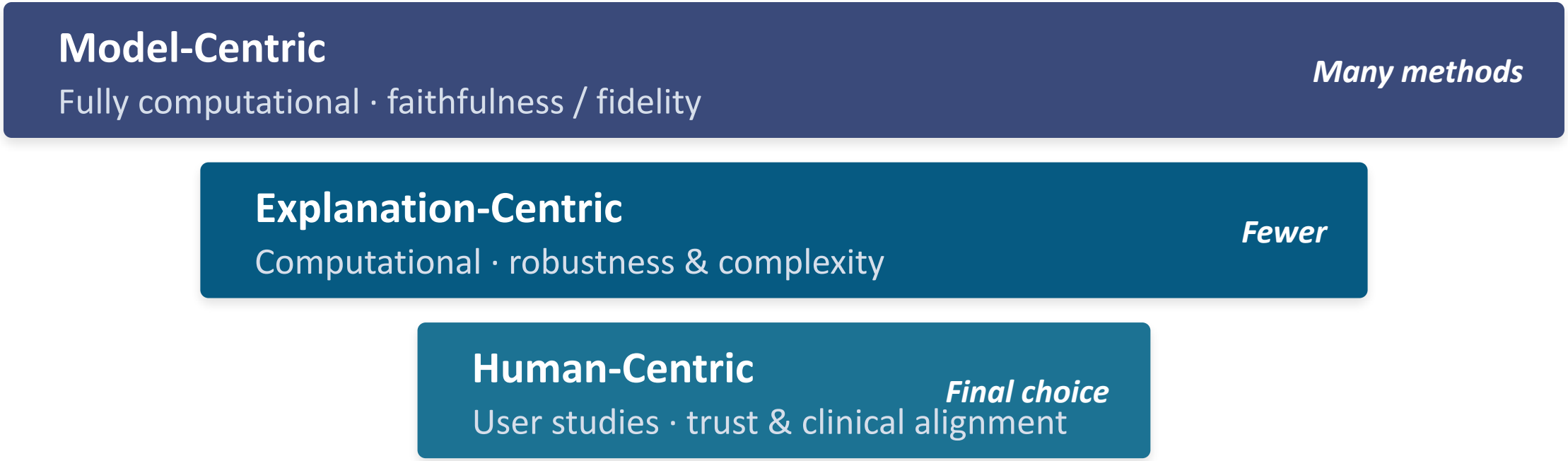
What each layer measures

Layer	Question it answers	Example metrics
Model-Centric	Does the explanation faithfully reflect the model?	Faithfulness, Fidelity, Monotonicity, Random Logit Test
Explanation-Centric	Is the explanation robust, stable and simple enough?	Max-Sensitivity, Local Lipschitz, Effective Complexity, stability metrics
Human-Centric	Do users find it useful, trustworthy & aligned?	Trust, Clinical Alignment, Expert Agreement, Over-Reliance

Synthesizes the categorizations of three major surveys (Chen et al. 2024 · Zhou et al. 2021 · Nauta et al. 2023) into one actionable hierarchy.

Each layer is a progressive filter

Layers are ordered by cost and by what they can prove. Cheap, fully-computational checks come first; expensive human studies come last — applied only to the few methods that survive.



Methods that fail a layer are dropped — they never reach costly human evaluation.

A gap in existing tools

Of 10 surveyed tools & frameworks, only **one (AutoXAI)** supports method **selection**. The rest only enable evaluation or comparison, leaving the user without guidance. Only **one (OpenHEXAI)** supports human-centric evaluation.

Open challenges

1. Context-specific guidance on which metrics fit which (model × task × domain)
2. Metric disagreement — metrics for the same property can conflict
3. Property trade-offs — strong faithfulness vs. strong robustness (Pareto)
4. Bridging computational metrics with human-centered evaluation
5. Real-world deployment in dynamic, pipeline-integrated systems

TAKEAWAYS

A structured path to method selection

A 3-layer pyramid Model-centric → explanation-centric → human-centric, evaluated bottom-up.

Progressive filtering Cheap computational checks first; costly human studies only for survivors.

Maps the open problems Metric disagreement, trade-offs, and bridging computation with human evaluation.

Thank you · Priscylla Silva, Victor H. Barella, Luis Gustavo Nonato

4th World Conference on eXplainable AI · Fortaleza, Brazil · July 2026

A Structured Guide to Selecting Feature Attribution Techniques

USP
GIA graphics, imaging, visualization and analytics
INSTITUTO FEDERAL Alagoas
GENIAR GRUPO DE PESQUISA EM EMPREENDEDORISMO, NETWORK, INTELIGÊNCIA ARTIFICIAL E ROBÓTICA
FAPESP SÃO PAULO RESEARCH FOUNDATION

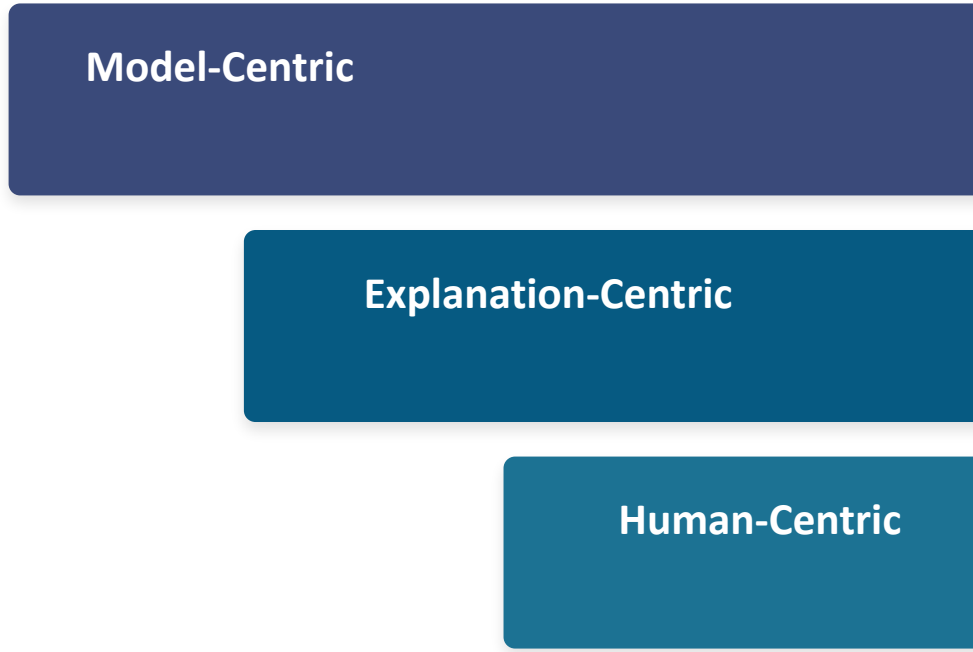
A heart-disease diagnostic system

A clinical team trains a deep neural network to predict heart disease and reaches high accuracy. Before deploying it to support cardiologists, they must explain its predictions and choose one attribution method to do so. We apply the pyramid to make that choice systematic.

High-stakes domain Healthcare — explanations must be trustworthy and clinically meaningful.

Accurate but opaque A deep neural network with strong accuracy, but a black-box decision process.

Validated by exaperts Explanations are shared with cardiologists and integrated into the system.

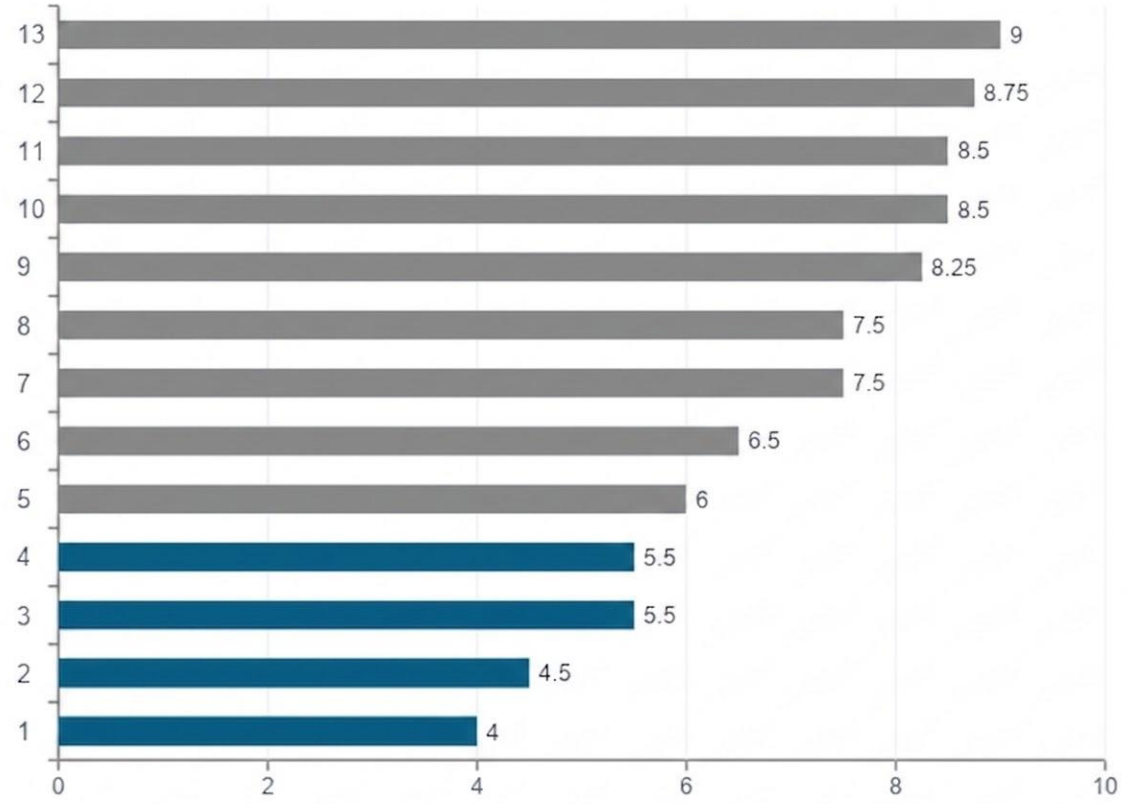


Model-Centric: 13 methods → 4

Thirteen attribution methods ranked by two fidelity metrics. The four with the strongest median ranks advance; the rest are dropped for poorly capturing the model.

Advance to Layer 2

- **Feature Ablation (4.0)**
- **SmoothGrad (4.5)**
- **Input × Gradient (5.5)**
- **Saliency (5.5)**

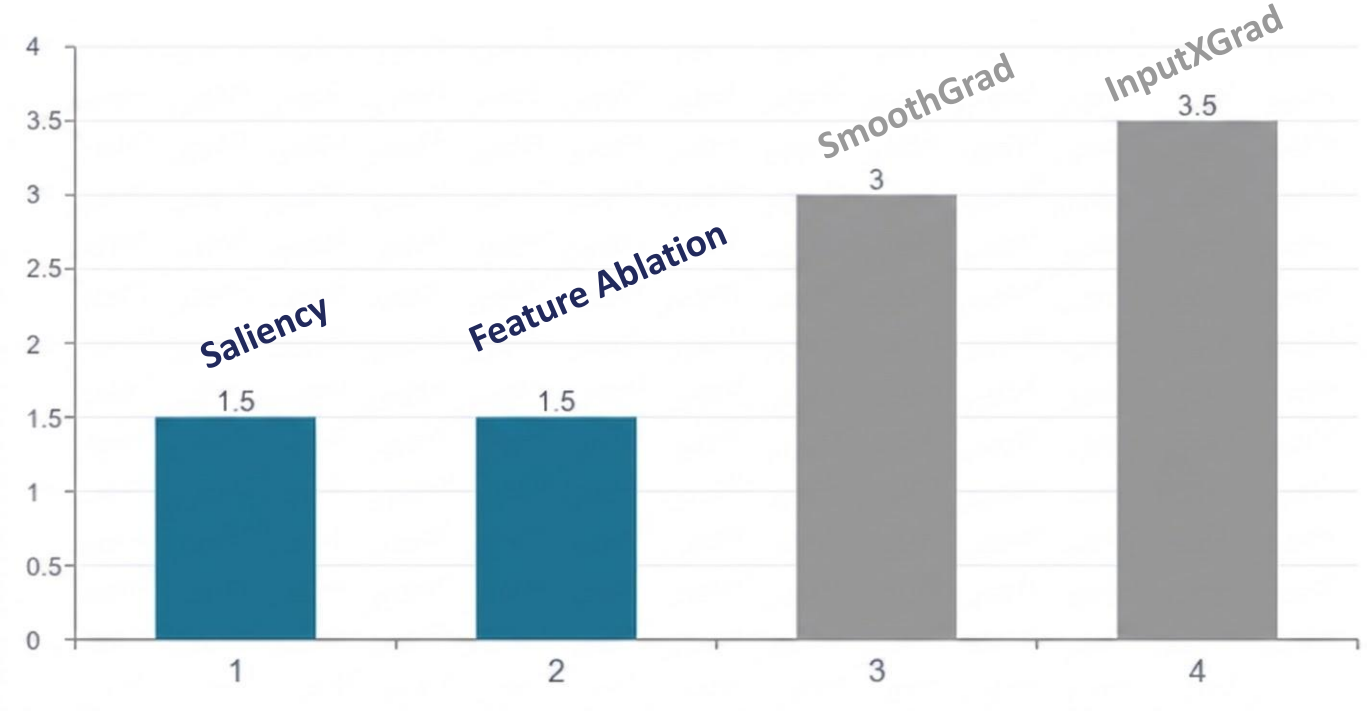


Overall median rank (lower = more faithful)

WORKED EXAMPLE · LAYER 2

Explanation-Centric: 4 methods → 2

The four survivors are tested on six robustness & complexity metrics. Feature Ablation and Saliency consistently rank best and advance.



Lower median rank = better across the six metrics

Survivors

Saliency

Most compact explanations (≈3–5 features) with high stability.

Feature Ablation

Best robustness to input perturbations.

WORKED EXAMPLE · LAYER 3

Human-Centric: 2 methods → 1

Three cardiologists evaluate the two finalists. Across every human-centered measure, Feature Ablation wins — and is selected for deployment.

Trust (1–5)

4.3

Feature Ablation

Saliency **3.8**

Clinical alignment (1–5)

4.1

Feature Ablation

Saliency **3.6**

Expert agreement

87%

Feature Ablation

Saliency **73%**

Result: 13 candidates systematically narrowed to 1 (Feature Ablation) for deployment.