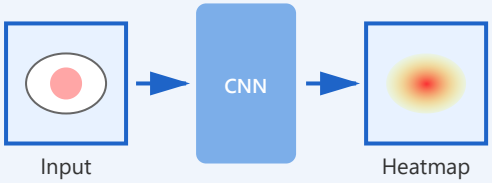
Explainable AI (XAI) - Visual Guide

1. Saliency-Based Methods

Grad-CAM



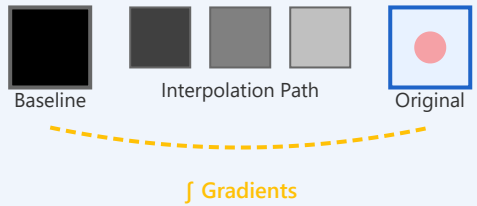
✓ Fast and class-discriminative

✓ Works with any CNN architecture

✗ Low spatial resolution

Example: Highlights lung regions with pneumonia

Integrated Gradients



✓ Theoretically grounded (completeness)

✓ Precise pixel-level attribution

✗ Computationally expensive

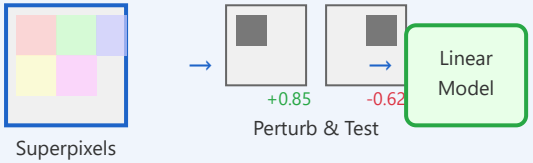
Example: Identifies exact pixels contributing to tumor classification

infiltrates in chest X-rays

Radiology Pathology

Oncology Diagnosis

🎯 LIME



✓ Model-agnostic approach

✓ Interpretable local explanations

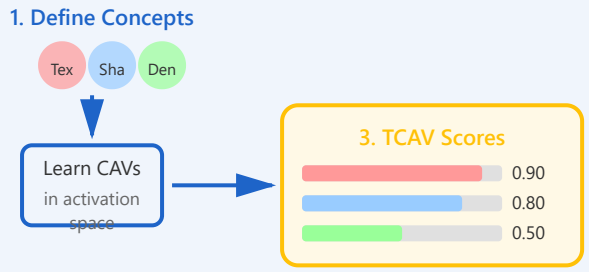
✗ Can be unstable

💡 *Example: Explains which image regions support/oppose a diagnosis*

Black-box Any model

2. Concept Attribution Methods

TCAV



✓ Uses clinically meaningful concepts

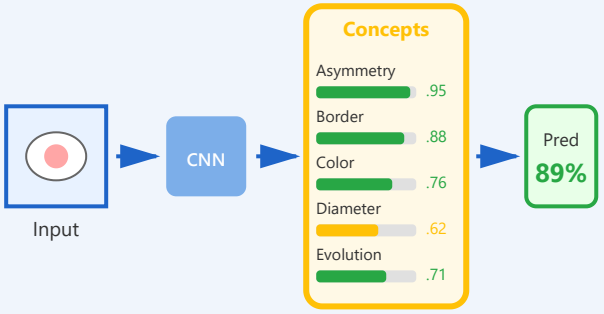
✓ Quantifies concept importance

✗ Requires concept example datasets

💡 *Example: "Model is 90% sensitive to nodular texture in lung CT scans"*

Clinical Concepts

Concept Bottleneck Models



✓ Built-in interpretability

✓ Transparent decision pathway

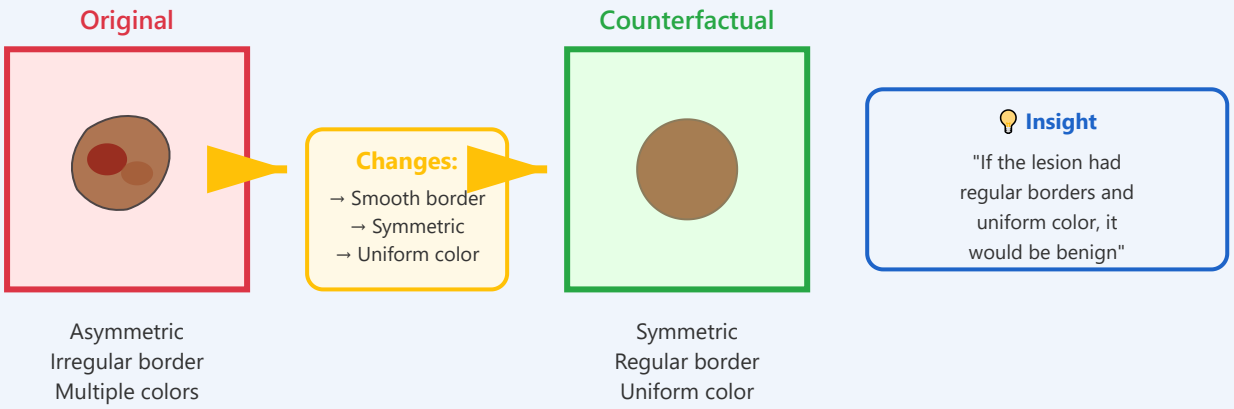
✗ May sacrifice some accuracy

💡 *Example: Melanoma detection using ABCDE criteria (visible intermediate layer)*

Dermatology ABCDE

3. Counterfactual Explanations

What-If Analysis



✓ Shows minimal changes needed to flip prediction

✓ Actionable insights for clinicians

✓ Helps understand decision boundaries

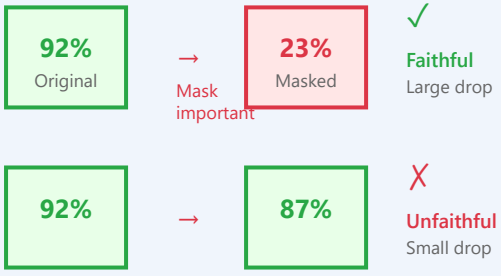
✗ May suggest unrealistic changes

Example: "If border irregularity decreased from 0.85 to 0.45, prediction would flip to benign"

Borderline What-if Actionable

4. Trust Building & Validation

✅ Faithfulness Testing



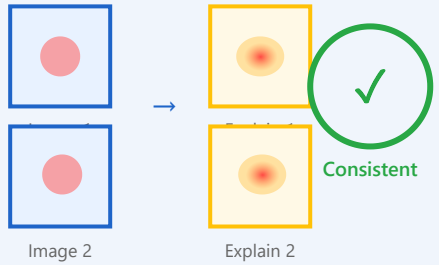
✓ Measures if explanations reflect actual model behavior

✓ Deletion/insertion tests

✗ Masking artifacts can affect results

Validation Metrics

🎯 Consistency Check



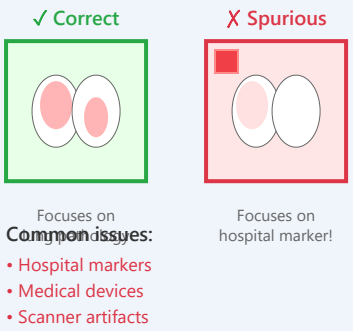
✓ Similar inputs → similar explanations

✓ Stable under small perturbations

✗ Some methods inherently unstable (LIME)

Stability Trust

⚠ Spurious Detection



✓ Identifies unreliable model reasoning

✓ Critical for safety

✗ Requires expert validation

💡 *Real case: Pneumonia model achieved 95% but focused on portable X-ray markers!*

Safety Validation

5. Interpretability Needs by Stakeholder

👨‍⚕️ Clinicians

Need:

- Case-specific explanations
- Clinical feature attribution
- Quick, actionable insights
- Uncertainty quantification

"Why is this lesion melanoma? Which features are most concerning?"

🧑 Patients

Need:

- Plain language
- What it means for health
- Treatment implications
- Confidence in AI use

"What does this result mean for me? Can I get a second opinion?"

⚖️ Regulators

Need:

- Auditability
- Bias detection
- Compliance (FDA, GDPR)
- Accountability trails

"Can we audit this decision if legally challenged?"

💻 AI Developers

Need:

- Model debugging
- Feature importance
- Failure mode detection
- Performance metrics

"Is the model using appropriate features? Where does it fail?"

Key Takeaways

Explainable AI: Essential for Medical AI Deployment

Saliency Methods

Visual attribution
Grad-CAM, LIME

Concept Attribution

Clinical concepts
TCAV, CBMs

Counterfactuals

What-if analysis
Actionable insights

Essential Requirements:

✓ Trust & Validation through faithfulness and consistency

✓ Stakeholder Alignment meeting diverse needs

✓ Safety through spurious correlation detection

✓ Regulatory Compliance (FDA, GDPR, EU AI Act)

"XAI enables clinicians to understand, validate, and appropriately rely on AI while maintaining safety, accountability, and trust"