Lecture 15 - Contents

An overview of the main sections in this lecture.

Part 1
Attribution and Visualization Methods

Part 2
Counterfactuals and Explanation Interfaces

Part 3
Communicating Uncertainty and Trade-offs

Hands-on
Interpretability Toolkit

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 15:

Explainable Medical AI: Building Trust Through Transparency

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com

Lecture Contents

Part 1: Attention-Based Interpretability

Part 2: Clinical Explanation Generation

Part 3: Clinical User Requirements

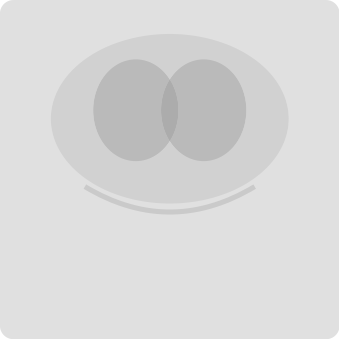
Part 1/3:

Attention-Based Interpretability

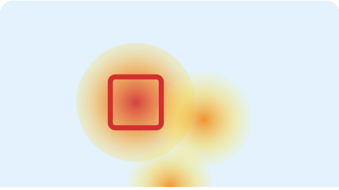
1. Attention Visualization
2. Layer-wise Relevance Propagation
3. Gradient-Based Attribution
4. Integrated Gradients
5. SHAP for Medical Applications
6. LIME for Clinical Text
7. Concept Activation Vectors

Attention Visualization

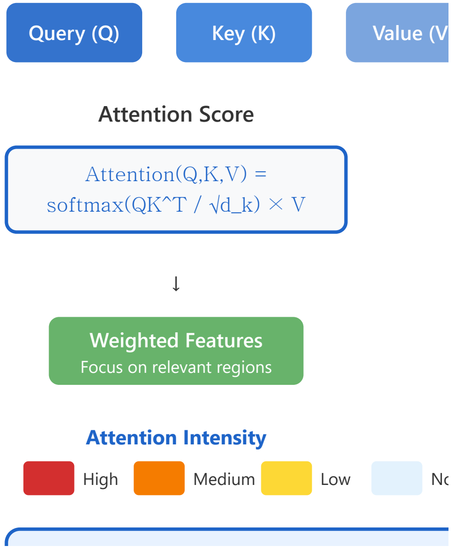
Medical Image (X-ray)



Attention Heatmap



Attention Mechanism



$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d\_k}) \times V$$

Attention Intensity

🟥 High 🟧 Medium 🟨 Low ⬜ Nc

🧠 Attention Mechanism

Neural network component showing where the model focuses when making predictions

🎨 Heatmap Visualization

Color-coded maps indicating attention weights across different input regions

🥼 Medical Image Analysis

Highlighting relevant regions in X-rays, CT scans, and MRI images

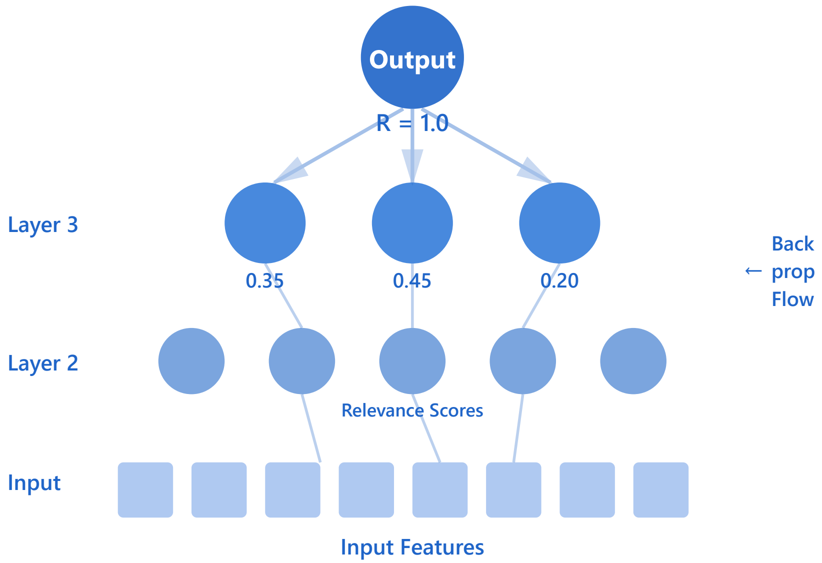
⚕ Clinical Application

Identifying pathological features automatically and visualizing model focus areas

💡 Key Benefit

Allows clinicians to verify that AI models are focusing on medically relevant features

Layer-wise Relevance Propagation (LRP)



🔄 LRP Principle

Backpropagating relevance scores from output to input layers

📊 Layer Contribution

Measuring each layer's contribution to the final prediction

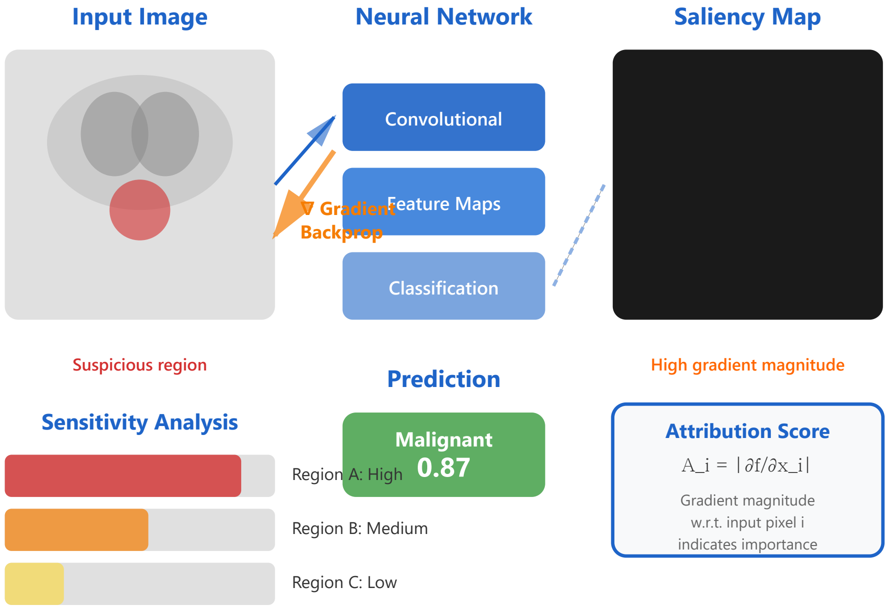
⚖️ Conservation Property

Total relevance remains constant through all layers

🥼 Medical Imaging

Identifying diagnostic features in medical images

Gradient-Based Attribution



Input Image

Neural Network

Saliency Map

Convolutional

Feature Maps

Classification

∇ Gradient Backprop

Suspicious region

High gradient magnitude

Prediction

Malignant 0.87

Sensitivity Analysis

Region A: High

Region B: Medium

Region C: Low

Attribution Score

$A_i = |\partial f/\partial x_i|$

Gradient magnitude w.r.t. input pixel i indicates importance

📈 Gradient Analysis

Using gradients to measure input feature importance

🎯 Sensitivity Mapping

Showing which inputs affect the output most significantly

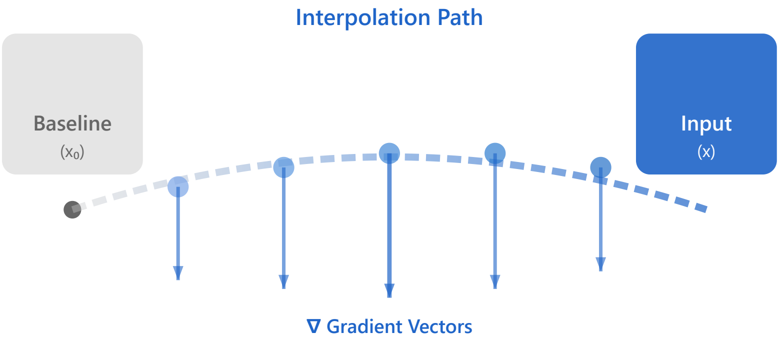
🗺️ Saliency Maps

Visualizing important pixels and regions in images

⚕️ Clinical Interpretation

Understanding what factors drive model decisions

Integrated Gradients



Attribution Formula

$$\mathrm{IG}(x) = (x - x_0) \times \int_{\alpha=0}^{\alpha=1} \nabla F(x_0 + \alpha(x - x_0))\, d\alpha$$

Sum of gradients along path

∫ Path Integration

Integrating gradients along interpolation path from baseline

🔍 Baseline Comparison

Comparing with neutral reference input (e.g., black image)

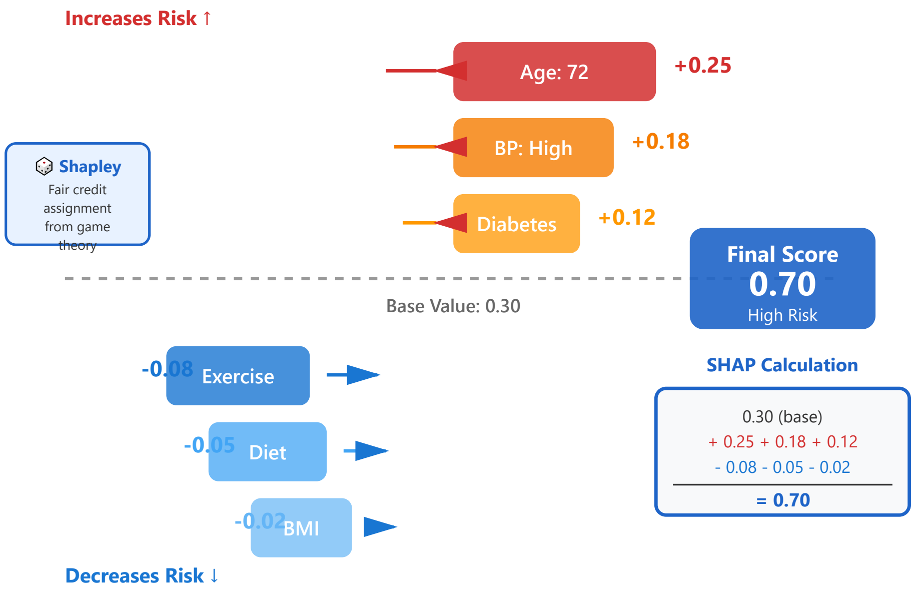
✓ Axiom Satisfaction

Meeting sensitivity and implementation invariance axioms

💪 Robust Attribution

More stable and reliable than simple gradient methods

SHAP Values for Medical Predictions



🎲 Shapley Values

Game-theoretic approach to measuring feature importance

➕ Additive Explanations

Consistent and locally accurate explanations

📊 Feature Contribution

Quantifying each feature's impact on the prediction

🩺 Clinical Decision Support

Explaining risk scores, diagnoses, and treatment recommendations

LIME for Clinical Text

🔍 Local Approximation

Explaining individual predictions with simple interpretable models

🔀 Text Perturbation

Systematically modifying input text to test importance

📝 Word Importance

Highlighting influential words and phrases in clinical notes

📋 EHR Text Analysis

Explaining decisions on electronic health record text data

Concept Activation Vectors (CAV)

💡 High-level Concepts

Testing for abstract medical concepts in neural networks

➡️ Direction Vectors

Finding concept directions in model activation space

🧪 Sensitivity Testing

Measuring model response to specific medical concepts

🔬 Medical Concepts

Detecting learned patterns like 'inflammation' or 'tumor characteristics'

Part 2/3:

Clinical Explanation Generation

1. Natural Language Explanations
2. Counterfactual Generation
3. Decision Path Visualization
4. Uncertainty Communication
5. Evidence Highlighting
6. Contrastive Explanations

Natural Language Explanations

💬 Text Generation

Generating human-readable explanations for predictions

🗣️ Clinical Language

Using medical terminology appropriate for healthcare professionals

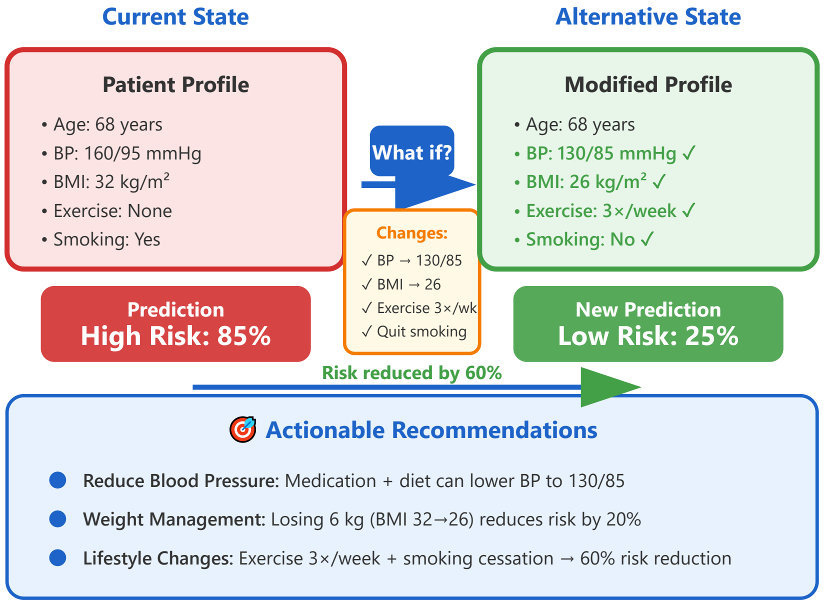
🔗 Multi-modal Integration

Combining image analysis with textual explanations

📖 Explanation Quality

Ensuring accuracy, completeness, and clinical relevance

Counterfactual Explanations



Current State

Patient Profile

- Age: 68 years
- BP: 160/95 mmHg
- BMI: 32 kg/m²
- Exercise: None
- Smoking: Yes

Prediction
High Risk: 85%

What if?

Changes:
✓ BP → 130/85
✓ BMI → 26
✓ Exercise 3×/wk
✓ Quit smoking

Alternative State

Modified Profile

- Age: 68 years
- BP: 130/85 mmHg ✓
- BMI: 26 kg/m² ✓
- Exercise: 3×/week ✓
- Smoking: No ✓

New Prediction
Low Risk: 25%

Risk reduced by 60%

🎯 Actionable Recommendations

- **Reduce Blood Pressure:** Medication + diet can lower BP to 130/85
- **Weight Management:** Losing 6 kg (BMI 32→26) reduces risk by 20%
- **Lifestyle Changes:** Exercise 3×/week + smoking cessation → 60% risk reduction

🔄 What-If Scenarios

Showing minimal changes needed to alter the prediction

🎯 Actionable Insights

Identifying modifiable factors that influence outcomes

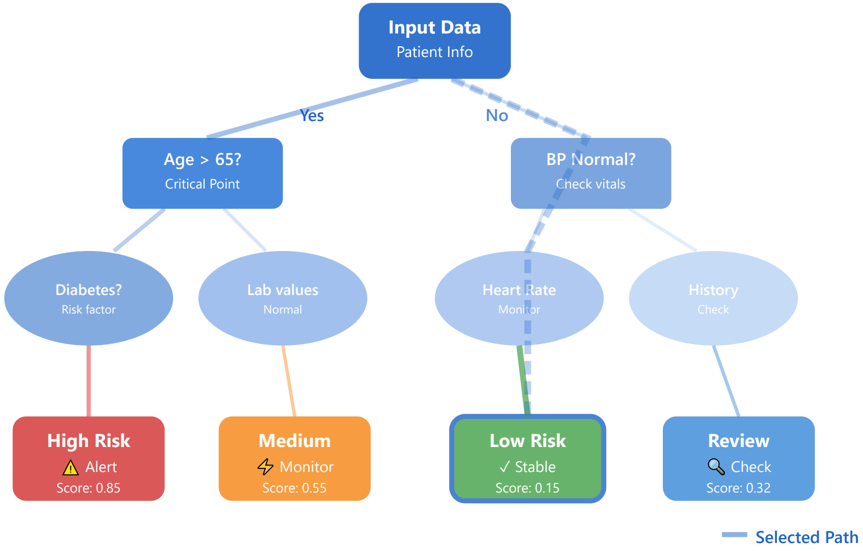
📊 Clinical Utility

Helping clinicians understand treatment alternatives

⚖️ Minimal Changes

Finding smallest modifications for desired outcome changes

Decision Path Visualization



🌳 Decision Trees

Visualizing hierarchical decision-making processes

🔍 Path Tracing

Following the model's reasoning from input to output

📍 Critical Points

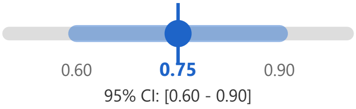
Identifying key decision points in the prediction process

🗺 Visual Navigation

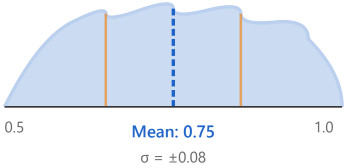
Interactive exploration of model decision logic

Uncertainty Communication

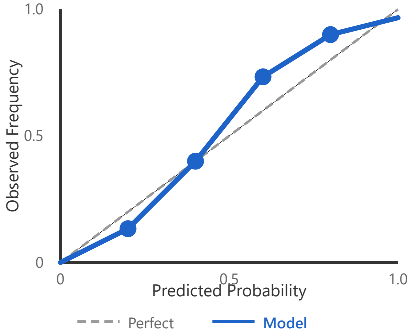
Confidence Interval



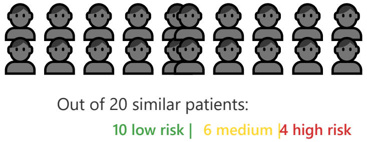
Probability Distribution



Calibration Curve



Risk Visualization



📊 Confidence Intervals

Quantifying prediction uncertainty with statistical bounds

🎲 Probabilistic Outputs

Expressing predictions as probability distributions

⚠ Risk Communication

Clearly conveying model confidence to clinicians

🎯 Calibrated Predictions

Ensuring predicted probabilities match actual frequencies

Evidence Highlighting

✨ Feature Marking

Highlighting relevant features that support the prediction

🔗 Source Attribution

Linking predictions to specific data sources

📚 Evidence Ranking

Ordering supporting evidence by importance

🔍 Traceability

Enabling verification of model reasoning process

Contrastive Explanations

⚖️ Comparison Analysis

Explaining why A rather than B was predicted

🔄 Differential Features

Identifying key differences between alternative outcomes

🎯 Foil Cases

Using contrasting examples to clarify decisions

💡 Enhanced Understanding

Improving comprehension through comparative reasoning

Part 3/3:

Clinical User Requirements

1. Physician Interpretability Needs
2. Patient-Facing Explanations
3. Regulatory Documentation
4. Audit Trail Generation
5. Trust Calibration
6. Error Analysis & Debugging

Physician Interpretability Needs

🧑‍⚕️ Clinical Workflow

Integrating explanations into existing medical workflows

⏱️ Time Efficiency

Providing quick, actionable insights without overwhelming detail

🎓 Medical Expertise

Matching explanation complexity to physician knowledge level

✅ Trust Building

Fostering appropriate trust through transparent AI

Patient-Facing Explanations

👥 Accessible Language

Using simple, non-technical terms for patient comprehension

🎨 Visual Communication

Employing graphics and diagrams for better understanding

💬 Empathy & Support

Providing explanations with emotional sensitivity

📖 Health Literacy

Adapting to varying levels of medical knowledge

Regulatory Documentation Requirements

📋 FDA Compliance

Meeting regulatory requirements for AI/ML medical devices

📝 Documentation Standards

Maintaining comprehensive model development records

🔍 Validation Evidence

Providing interpretability as part of validation process

✓ Approval Process

Supporting regulatory submissions with explainability data

Audit Trail Generation

📊 Decision Logging

Recording all model inputs, outputs, and reasoning steps

🕐 Temporal Tracking

Timestamping predictions and model versions

🔒 Immutable Records

Creating tamper-proof logs for legal compliance

🔍 Retrospective Analysis

Enabling investigation of past predictions

Trust Calibration

⚖️ Appropriate Trust

Balancing between over-reliance and under-utilization

📈 Performance Awareness

Communicating model strengths and limitations clearly

⚠️ Failure Cases

Highlighting scenarios where model may be unreliable

🎯 Confidence Alignment

Ensuring user trust matches actual model performance

Error Analysis Tools

🔍 Failure Detection

Identifying systematic errors and edge cases

📊 Error Classification

Categorizing different types of prediction failures

🎯 Root Cause Analysis

Understanding why specific errors occur

🔧 Improvement Insights

Using error patterns to guide model refinement

Debugging Interfaces for Developers

🖥️ Interactive Tools

Visual interfaces for exploring model behavior

🔬 Layer Inspection

Examining activations and weights at each layer

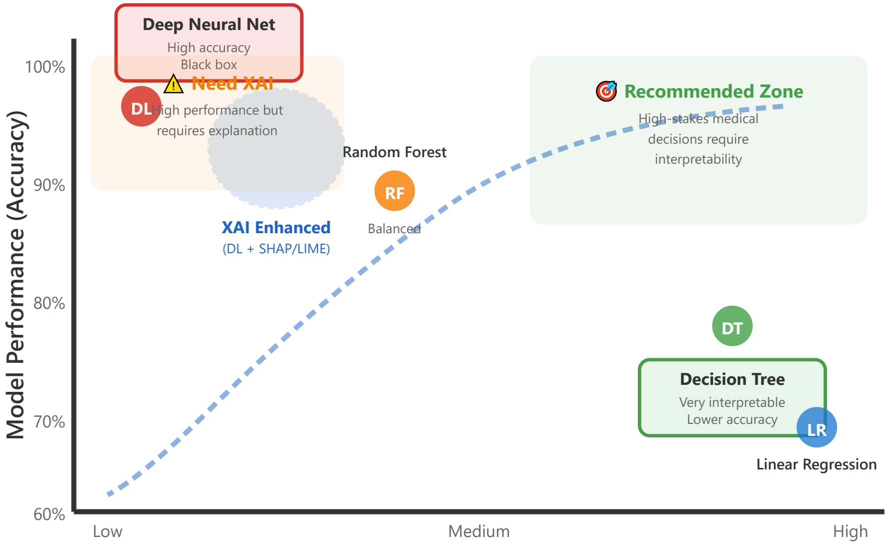
📊 Performance Metrics

Real-time monitoring of model performance indicators

🐛 Bug Identification

Tools for detecting and fixing model issues

Performance vs. Interpretability Trade-off



Context matters:ICU predictions → Need interpretability | Image screening → Accept some opacity | Risk scoring → Balance botl

⚖️ Accuracy vs. Transparency

Balancing model performance with explainability

🧩 Model Complexity

Simpler models are more interpretable but may be less accurate

🎯 Application Context

High-stakes medical decisions may require more interpretability

💡 Hybrid Approaches

Combining powerful models with post-hoc explanations

Case Study: ICU Mortality Predictions

🏥 Clinical Context

Predicting patient mortality risk in intensive care units

📊 Feature Importance

Identifying vital signs and lab values driving predictions

⏱️ Temporal Patterns

Explaining how patient trajectory affects risk scores

🧑‍⚕️ Clinical Validation

Physicians reviewing and validating AI explanations

Hands-On: XAI Tools Implementation

🔧 SHAP Library

Python implementation: shap.TreeExplainer, shap.DeepExplainer

🎨 LIME Package

Text and image explainers: lime.lime_text, lime.lime_image

📊 Captum (PyTorch)

Integrated Gradients, GradCAM, and other attribution methods

💻 Practice Exercise

Implementing explanations for medical image classification

Future Research Directions in XAI

🔮 Multimodal Explanations

Combining imaging, text, and structured data explanations

🤖 Foundation Model XAI

Explaining large language and vision models in medicine

⚖️ Standardization

Developing common evaluation metrics for explainability

🏥 Clinical Integration

Seamlessly embedding XAI into electronic health records

Thank you

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com