Lecture 15 - Contents

An overview of the main sections in this lecture.

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

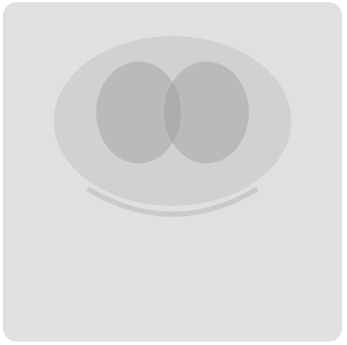
Part 1/3:

Attention-Based Interpretability

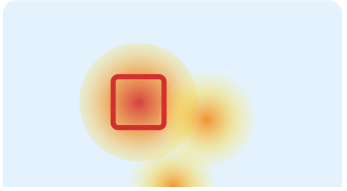
1. Attention Visualization
2. Layer-wise Relevance Propagation
3. Gradient-Based Attribution
4. Integrated Gradients
5. SHAP for Medical Applications
6. LIME for Clinical Text
7. Concept Activation Vectors

Attention Visualization

Medical Image (X-ray)



Attention Heatmap



Attention Mechanism

Query (Q) | Key (K) | Value (V

Attention Score

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

↓

Weighted Features
Focus on relevant regions

Attention Intensity

High | Medium | Low | No

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

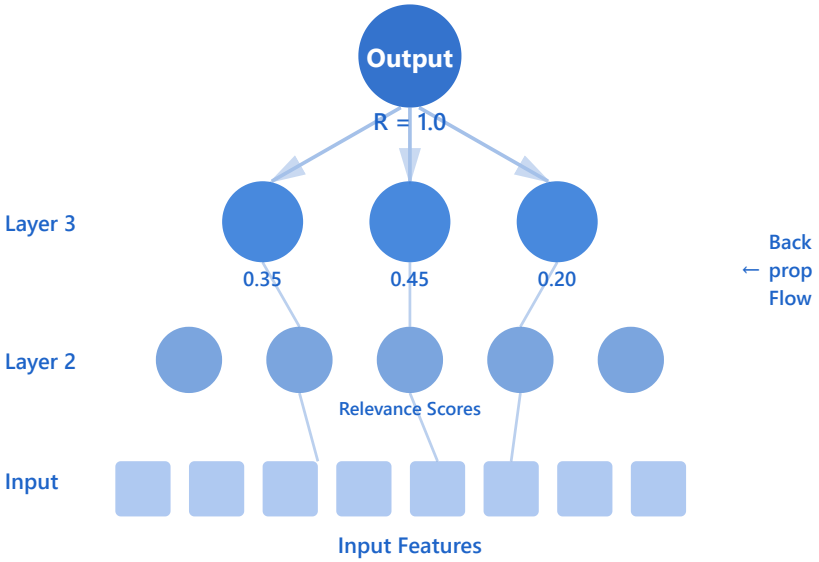
Clinical Application

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

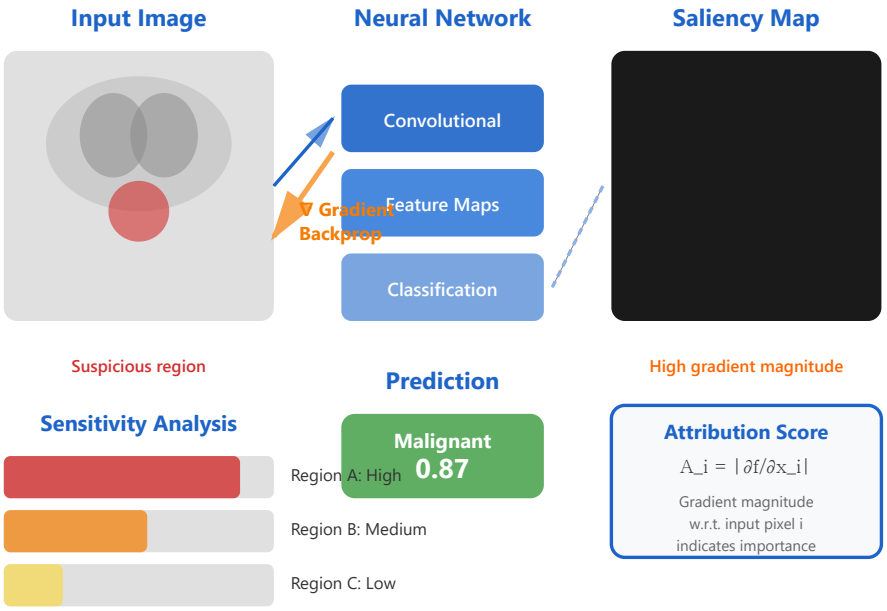
Conservation Property

Total relevance remains constant through all layers

Medical Imaging

Identifying diagnostic features in medical images

Gradient-Based Attribution



📈 Gradient Analysis

Using gradients to measure input feature importance

🎯 Sensitivity Mapping

Showing which inputs affect the output most significantly

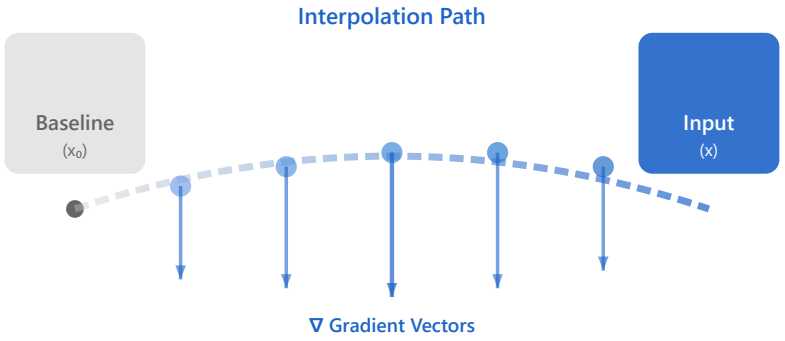
🗺️ Saliency Maps

Visualizing important pixels and regions in images

⚕️ Clinical Interpretation

Understanding what factors drive model decisions

Integrated Gradients



Attribution Formula

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

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)

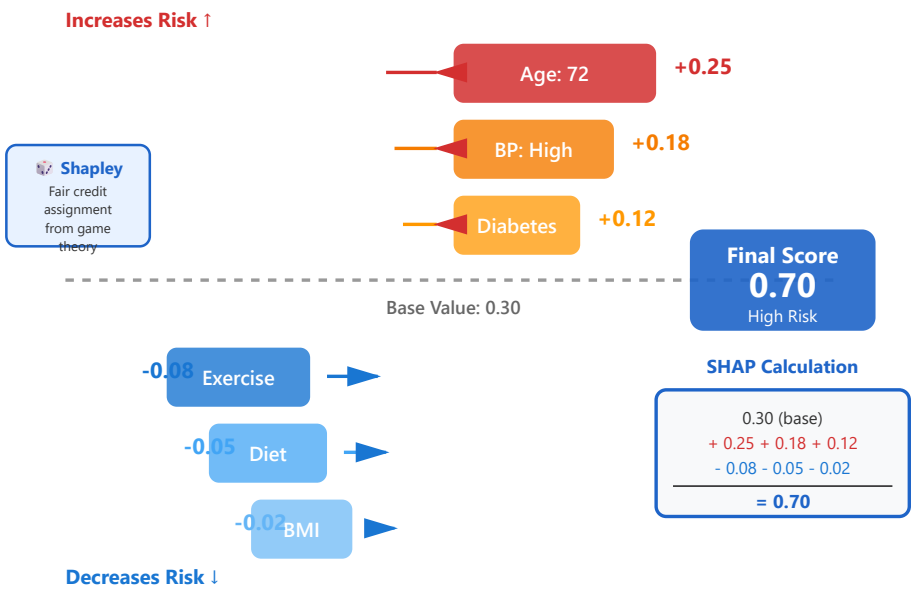
✓ Axiom Satisfaction

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

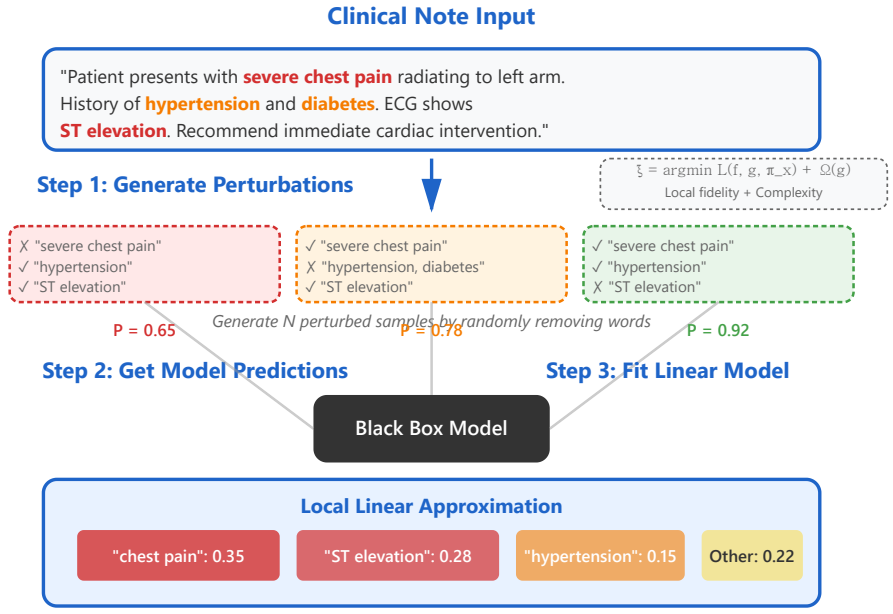
Feature Contribution

Quantifying each feature's impact on the prediction

Clinical Decision Support

Explaining risk scores, diagnoses, and treatment recommendations

LIME for Clinical Text



🔍 Local Approximation

LIME explains individual predictions by approximating the black box locally

🔀 Text Perturbation

Randomly removes words to create variations of the original text

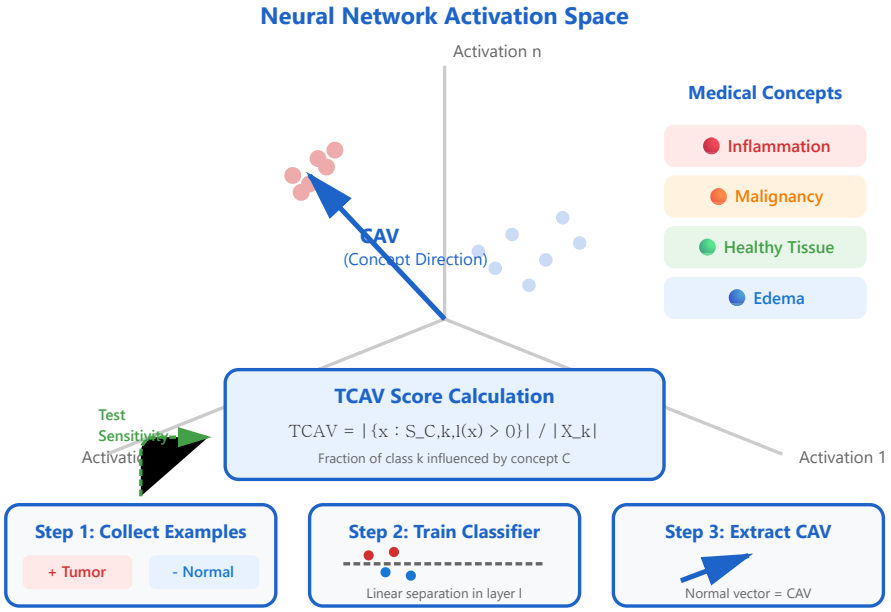
📊 Feature Importance

Linear model weights reveal word importance for the prediction

🏥 Clinical Application

Identifies key clinical terms driving diagnostic decisions

Concept Activation Vectors (CAV)



💡 High-level Concepts

CAVs test whether neural networks learn human-interpretable concepts

➡️ Direction Vectors

Concepts represented as directions in activation space

🧪 Sensitivity Testing

Measure how much a concept influences model predictions

🔬 Medical Validation

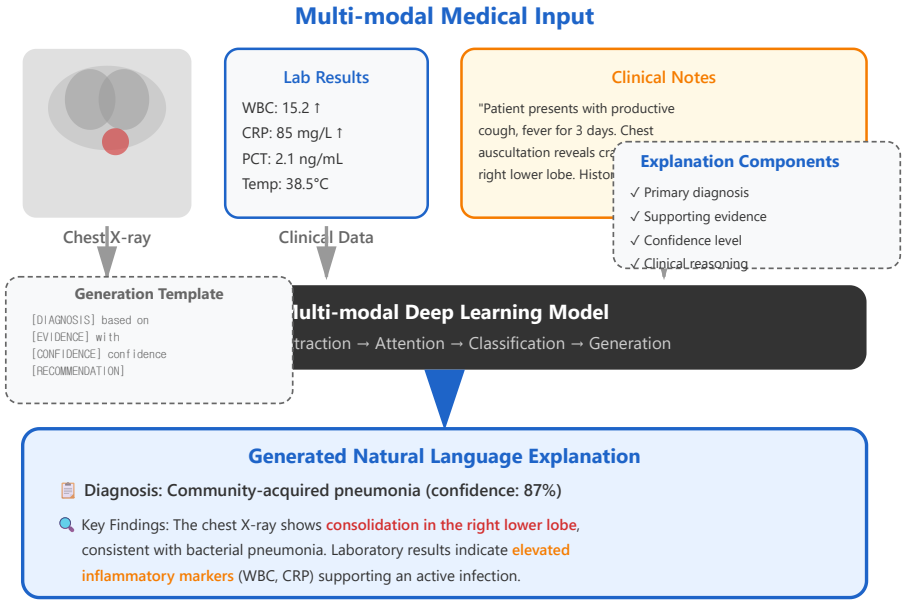
Verify if models use clinically relevant concepts

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

AI generates human-readable explanations automatically from model outputs

Clinical Language

Uses appropriate medical terminology for healthcare professionals

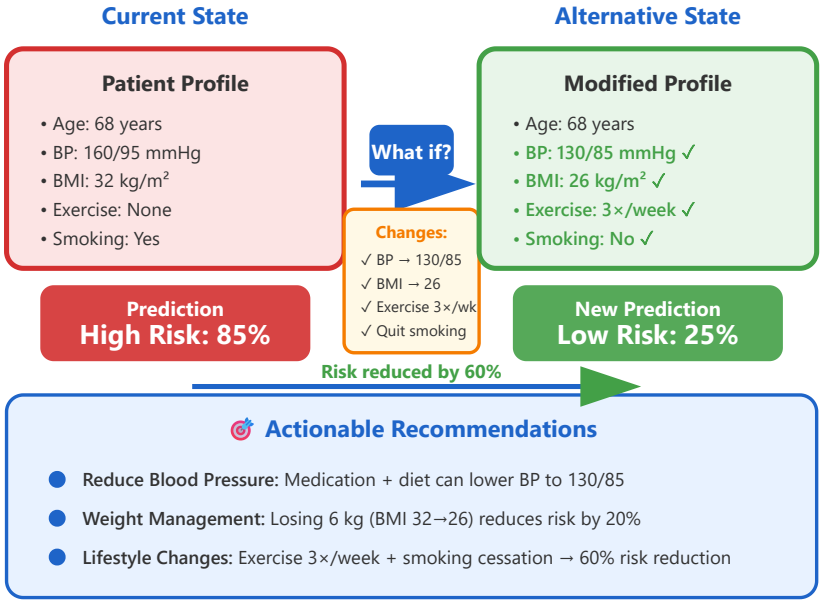
Multi-modal Integration

Combines insights from images, labs, and text into coherent explanations

Quality Assurance

Ensures accuracy, completeness, and clinical relevance of explanations

Counterfactual Explanations



What-If Scenarios

Showing minimal changes needed to alter the prediction

Actionable Insights

Identifying modifiable factors that influence outcomes

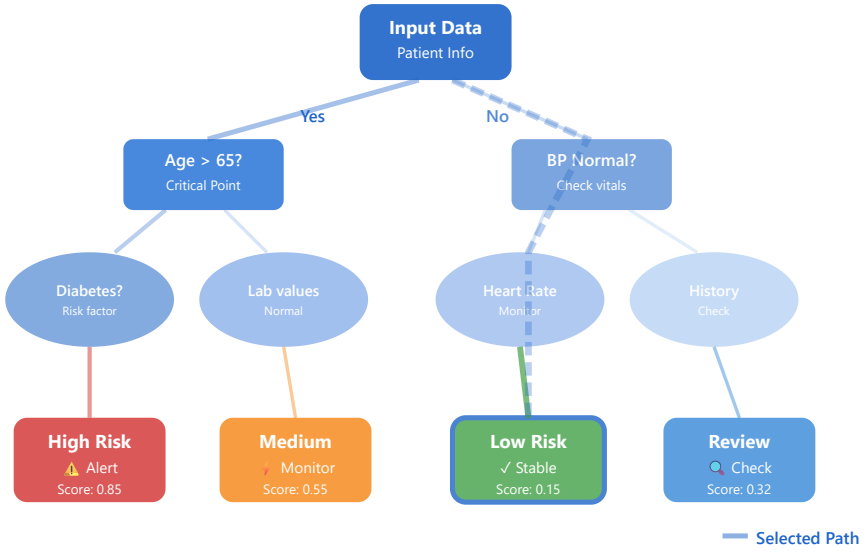
Clinical Utility

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

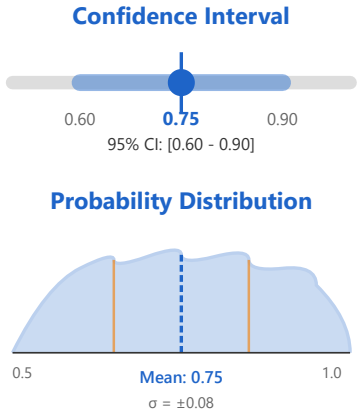
📍 Critical Points

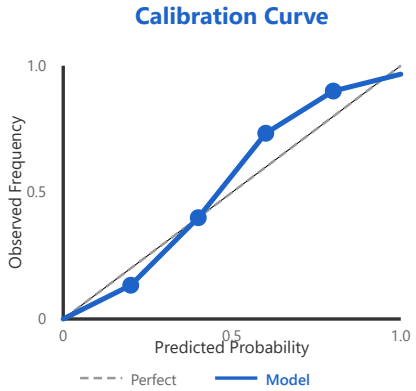
Identifying key decision points in the prediction process

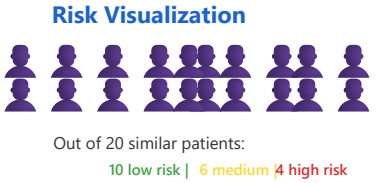
🗺️ Visual Navigation

Interactive exploration of model decision logic

Uncertainty Communication







📊 Confidence 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

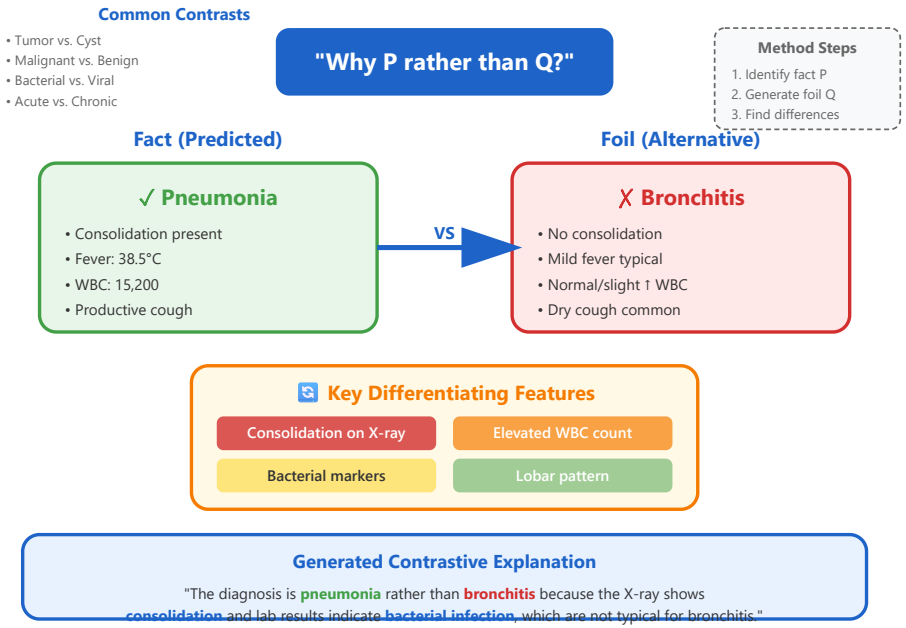
📚 Evidence Ranking

Ordering supporting evidence by importance

🔍 Traceability

Enabling verification of model reasoning process

Contrastive Explanations



⚖️ Comparison Analysis

Explains why the model chose one diagnosis over plausible alternatives

🔄 Differential Features

Identifies the specific features that distinguish between outcomes

🎯 Foil Generation

Creates meaningful alternative scenarios for comparison

💡 Enhanced Understanding

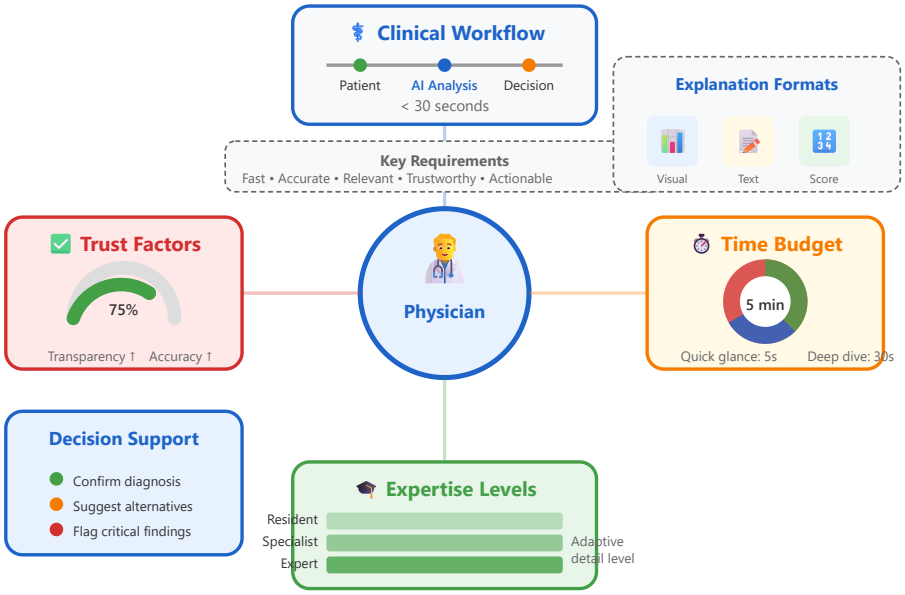
Improves comprehension by showing what the diagnosis is NOT

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 Integration

XAI must fit seamlessly into existing medical workflows without disruption

Time Constraints

Explanations must be digestible within 5-30 seconds during patient care

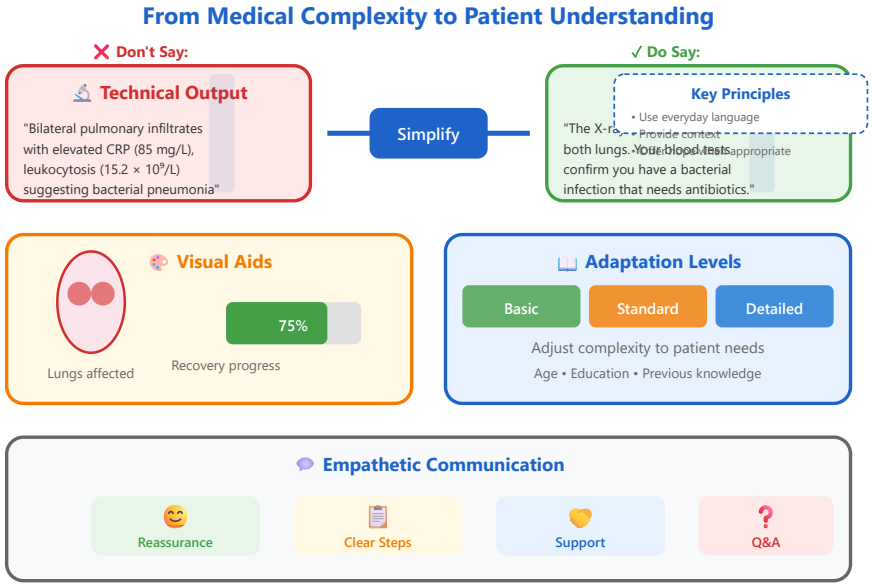
Adaptive Complexity

Adjust explanation detail based on physician's expertise and context

Trust Calibration

Build appropriate trust through transparency while avoiding over-reliance

Patient-Facing Explanations



Accessible Language

Translate medical jargon into everyday terms patients understand

Visual Communication

Use diagrams, charts, and illustrations to enhance understanding

Empathetic Approach

Deliver information with emotional sensitivity and support

Adaptive Complexity

Adjust explanation detail based on patient's health literacy level

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

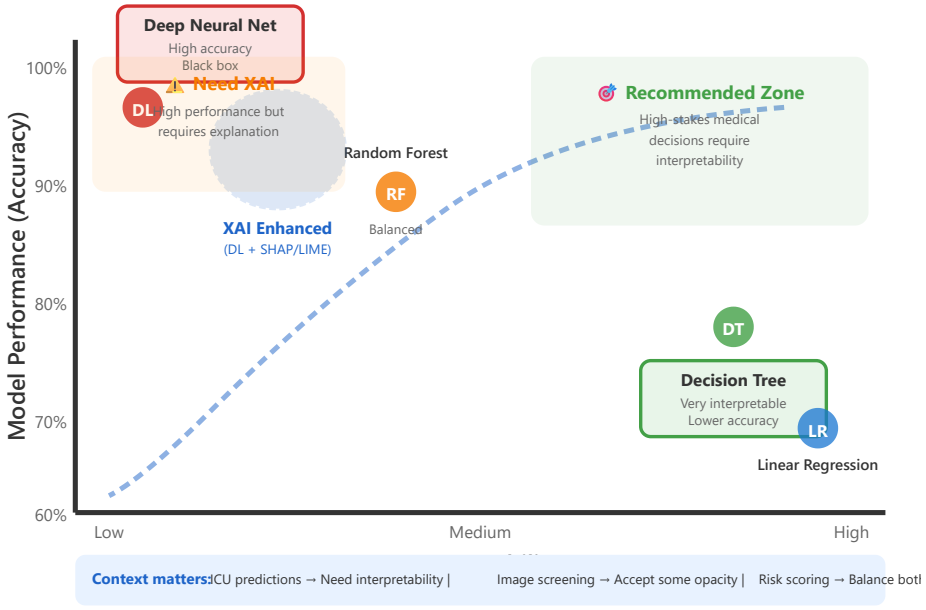
📊 Performance Metrics

Real-time monitoring of model performance indicators

🐛 Bug Identification

Tools for detecting and fixing model issues

Performance vs. Interpretability Trade-off



⚖️ 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