Explaining Black Box Predictions in Medical AI

Brian Huynh[2] Zoya Hasan[2] Joshua Lee[2] Mentor: Albert Hsiao [1,2]
bth001@ucsd.edu, zohasan@ucsd.edu, jdlee@ucsd.edu, a3hsiao@health.ucsd.edu
[1]Department of Radiology, University of California, San Diego; [2]University of California, San Diego

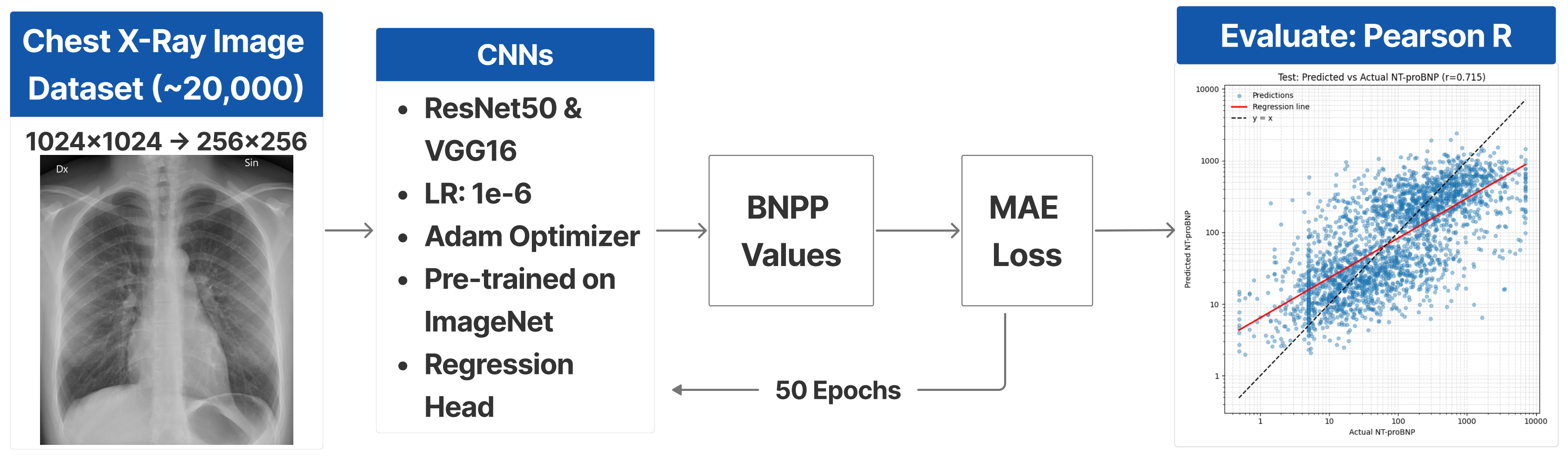
AiDA lab
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Introduction

- Pulmonary edema is a life-threatening condition characterized by fluid accumulation in the lungs, often caused by heart failure.
- Convolutional Neural Networks (CNNs) and Large Language Models (LLMs) show strong potential for detecting pulmonary edema from chest radiographs and radiology reports, respectively. However, their limited interpretability restricts clinical trust.
- We evaluate model-agnostic and model-specific methods on CNNs and LLMs to learn the features driving CNN and LLM predictions.

CNN (VGG16 & ResNet50) Development & Explainability

Imaging: We trained VGG16 and ResNet50 CNNs to predict a N-terminal pro-B-type natriuretic peptide (BNPP) biomarker value from grayscale chest X-rays. Final classification layers were replaced with a single linear output to extract continuous values.



Chest X-Ray Image Dataset (~20,000)
1024×1024 → 256×256

CNNs
- ResNet50 & VGG16
- LR: 1e-6
- Adam Optimizer
- Pre-trained on ImageNet
- Regression Head

BNPP Values

MAE Loss

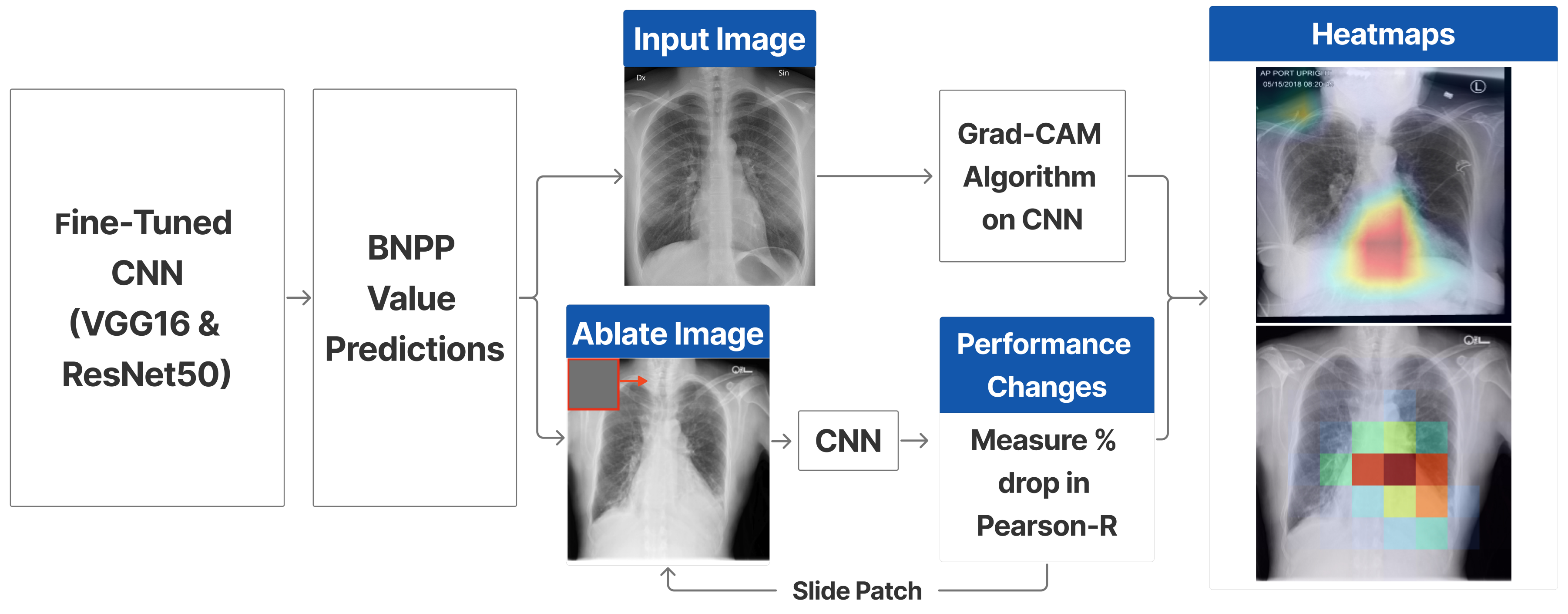
50 Epochs

Evaluate: Pearson R

Figure 1. CNN (VGG16 & ResNet50) Training Pipeline

Grad-CAM (Model-Specific) generates a heatmap by analyzing how sensitive the model's prediction is to patterns detected in its final convolutional layer, highlighting image regions that contribute most strongly to the output.

Image ablation (Model-Agnostic) systematically masks localized patches of the image and measures the resulting change in prediction, directly testing how important each region is.



Fine-Tuned CNN (VGG16 & ResNet50)

BNPP Value Predictions

Input Image

Grad-CAM Algorithm on CNN

Ablate Image

CNN

Performance Changes

Measure % drop in Pearson-R

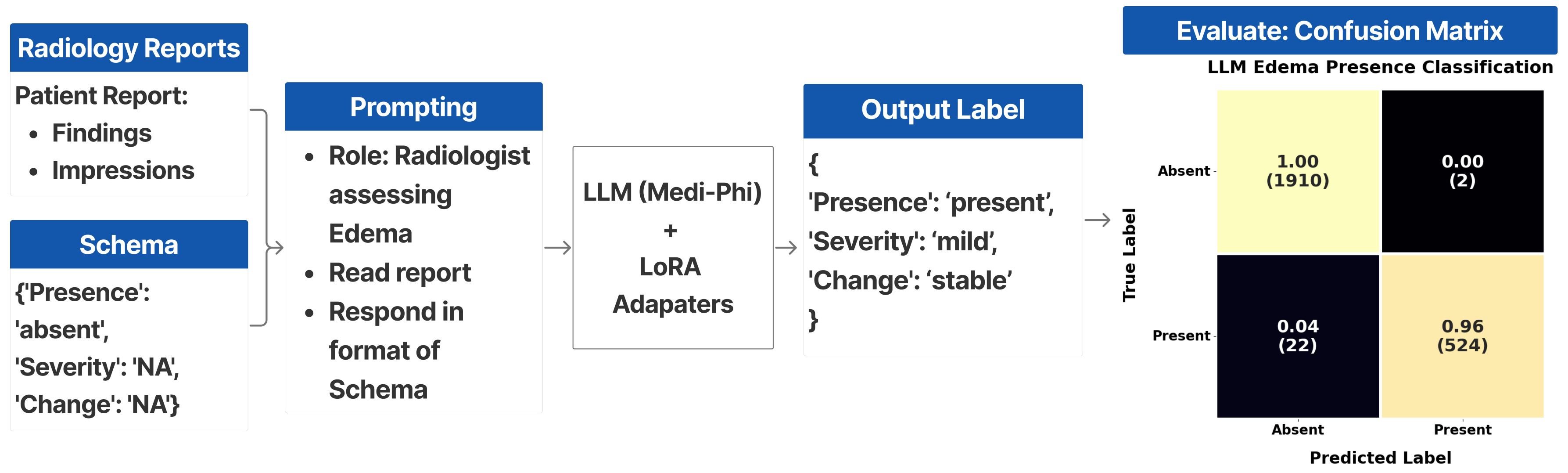
Slide Patch

Heatmaps

Figure 2. CNN Interpretability Workflow

LLM (MediPhi) Development & Explainability

Text: We fine-tuned Microsoft's MediPhi LLM using Low-Rank Adaptation (LoRA) to extract structured edema outcomes (presence and severity) from structured radiology reports. Base model weights were frozen, and only lightweight adapter layers were optimized.



Radiology Reports

Patient Report:
- Findings
- Impressions

Schema

{'Presence': 'absent', 'Severity': 'NA', 'Change': 'NA'}

Prompting
- Role: Radiologist assessing Edema
- Read report
- Respond in format of Schema

LLM (Medi-Phi) + LoRA Adapaters

Output Label

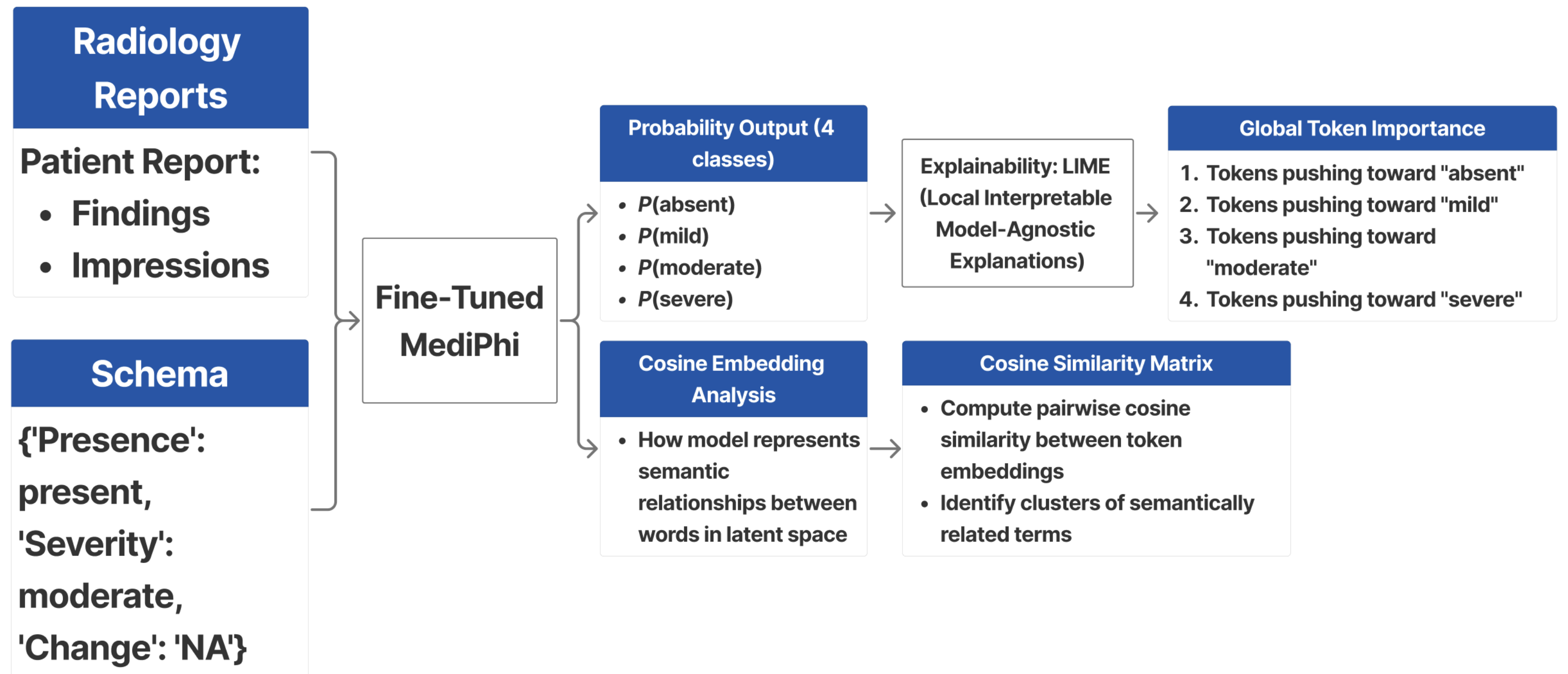
{ 'Presence': 'present', 'Severity': 'mild', 'Change': 'stable' }

Evaluate: Confusion Matrix

Figure 3. LLM Fine-tuning Pipeline

LIME Explainability (Model-Agnostic) Radiology reports are processed through a fine-tuned MediPhi model to generate four-class probability outputs (absent, mild, moderate, severe), and LIME (Local Interpretable Model-Agnostic Explanations) is applied to identify which tokens most influence each classification.

Cosine Embedding Analysis (Model-Specific) Cosine similarity is used to examine how the model internally groups clinically related words in its embedding space, revealing how terms like edema, congestion, and negation are semantically organized beyond direct prediction attribution.



Radiology Reports

Patient Report:
- Findings
- Impressions

Schema

{'Presence': present, 'Severity': moderate, 'Change': 'NA'}

Fine-Tuned MediPhi

Probability Output (4 classes)
- *P*(absent)
- *P*(mild)
- *P*(moderate)
- *P*(severe)

Explainability: LIME (Local Interpretable Model-Agnostic Explanations)

Global Token Importance
1. Tokens pushing toward "absent"
2. Tokens pushing toward "mild"
3. Tokens pushing toward "moderate"
4. Tokens pushing toward "severe"

Cosine Embedding Analysis
- How model represents semantic relationships between words in latent space

Cosine Similarity Matrix
- Compute pairwise cosine similarity between token embeddings
- Identify clusters of semantically related terms

Figure 4. LLM Explainability Workflow

Results

- **Image Ablation:** Both VGG16 and ResNet50 showed the largest performance drop when masking central chest regions, with minimal impact from masking image borders.
- **Grad-CAM:** VGG16 displayed widespread activation across the chest with attention to lung and cardiac regions, while ResNet50 showed more consistent and concentrated focus on the heart with additional emphasis in the lungs.

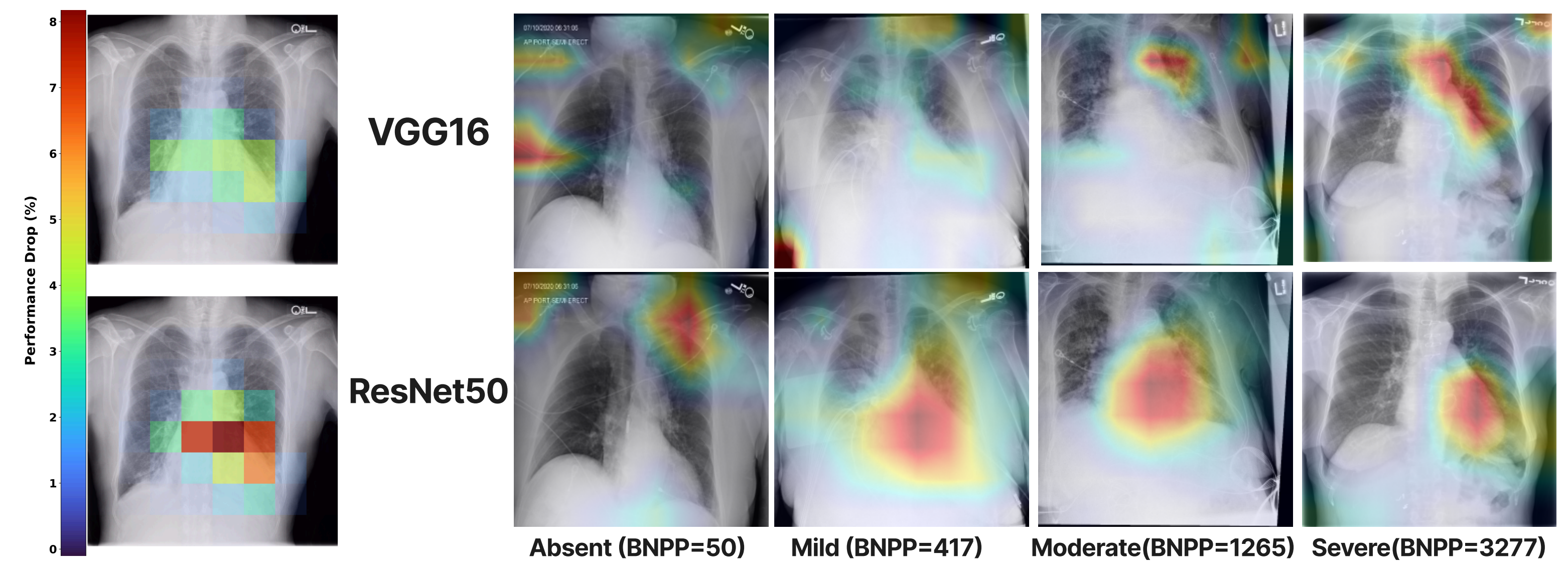


Figure 5. Left: Performance drop after regional image ablation. Right: Grad-CAM heatmaps across edema severity groups.

- **Global LIME Analysis:** Predictions align with clinically meaningful language. Higher severity is driven by terms indicating disease intensity and progression, while absent or lower severity is associated with stable or negating language.

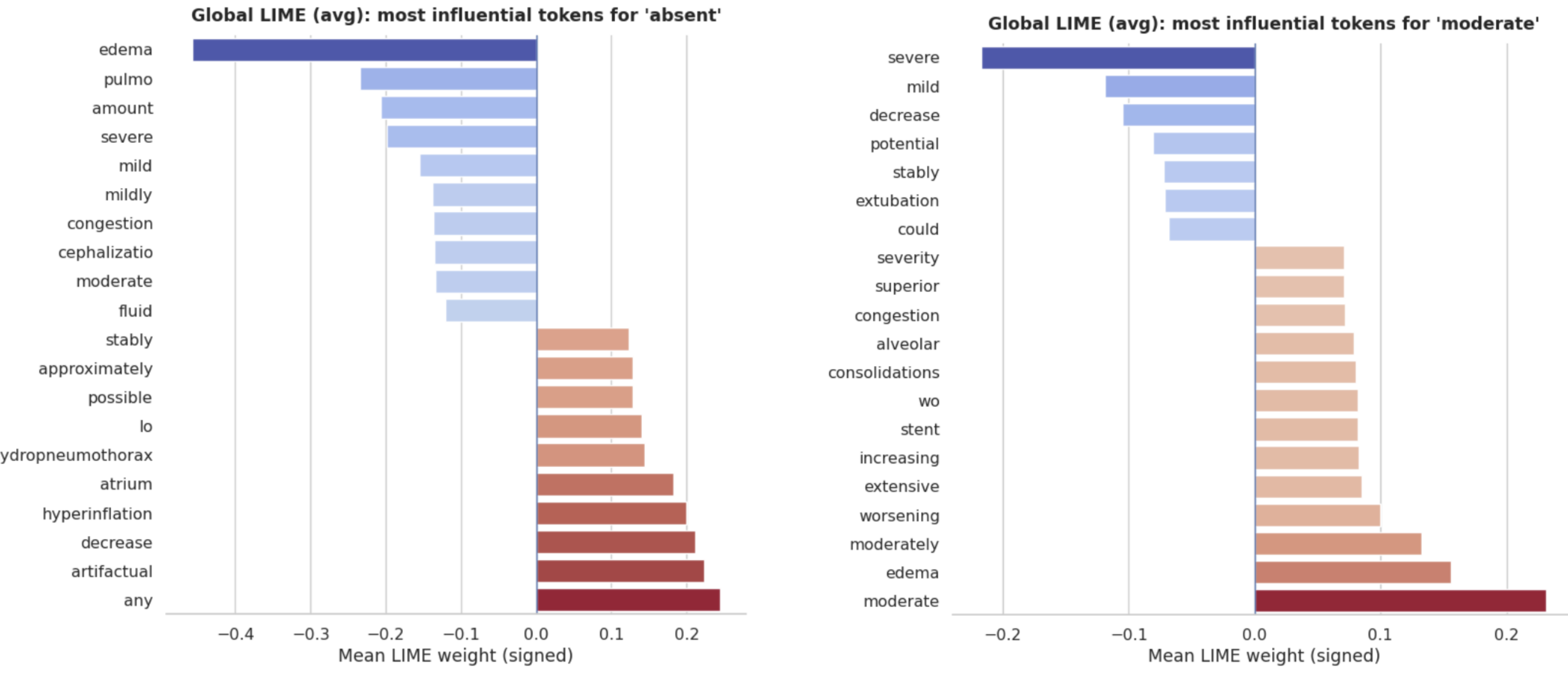
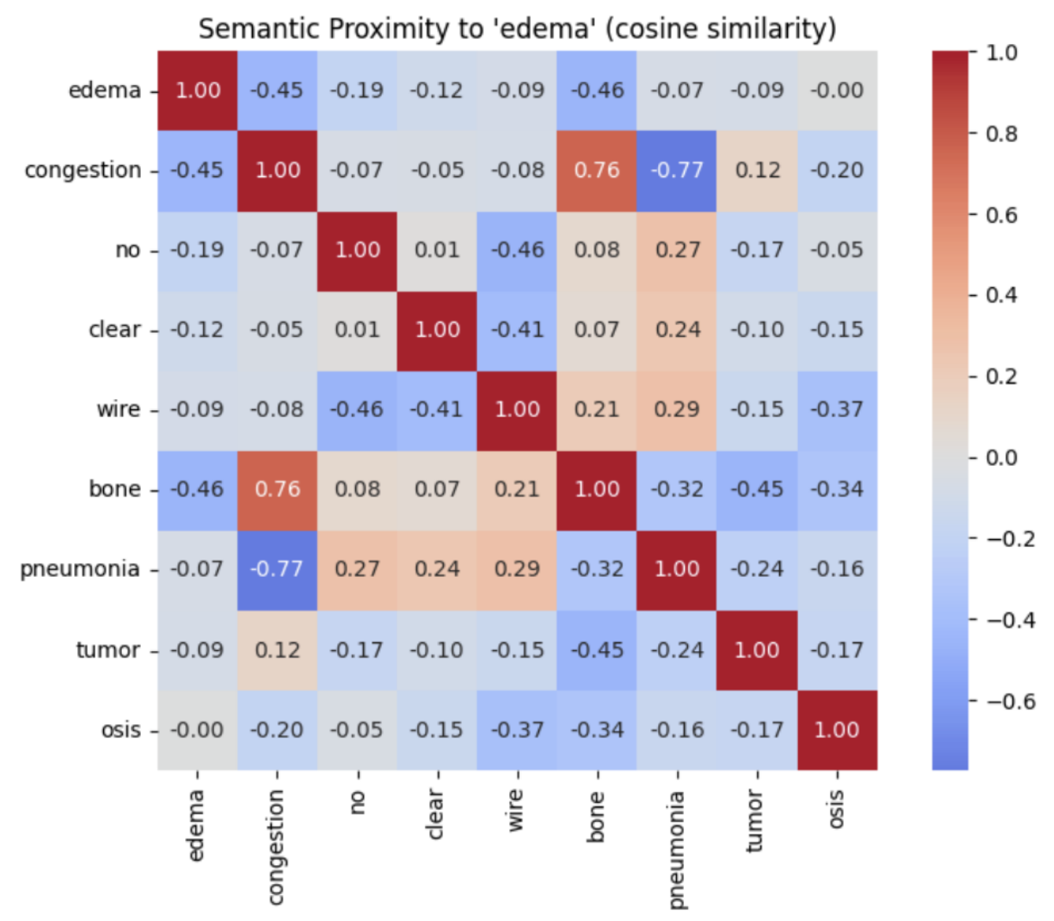


Figure 6. Global LIME token importance for selected edema classes. Positive weights support the class; negative weights oppose it.

- **Cosine Embedding Analysis:** The embedding space preserves meaningful clinical structure, with some related concepts clustering together while normal or negation-related language shifts in opposing directions.



	edema	congestion	no	clear	wire	bone	pneumonia	tumor	osis
edema	1.00	-0.45	-0.19	-0.12	-0.09	-0.46	-0.07	-0.09	-0.00
congestion	-0.45	1.00	-0.07	-0.05	-0.08	0.76	-0.77	0.12	-0.20
no	-0.19	-0.07	1.00	0.01	-0.46	0.08	0.27	-0.17	-0.05
clear	-0.12	-0.05	0.01	1.00	-0.41	0.07	0.24	-0.10	-0.15
wire	-0.09	-0.08	-0.46	-0.41	1.00	0.21	0.29	-0.15	-0.37
bone	-0.46	0.76	0.08	0.07	0.21	1.00	-0.32	-0.45	-0.34
pneumonia	-0.07	-0.77	0.27	0.24	0.29	-0.32	1.00	-0.24	-0.16
tumor	-0.09	0.12	-0.17	-0.10	-0.15	-0.45	-0.24	1.00	-0.17
osis	-0.00	-0.20	-0.05	-0.15	-0.37	-0.34	-0.16	-0.17	1.00

Figure 7. Cosine similarity among selected edema-related clinical terms. Red indicates alignment; blue indicates opposition.

Discussion & Conclusion

- CNN attention centers on cardiac and perihilar lung regions, highlighting key areas for pulmonary edema detection.
- ResNet50 exhibits more localized activation than VGG16, suggesting deeper networks capture more focused imaging features.
- MediPhi explanations indicate reliance on descriptive severity language, while embedding analysis reveals meaningful semantic relationships among clinically relevant terms.
- Combining model-agnostic and model-specific explanations reveals both what inputs drive predictions and how the model internally represents clinical features.