Explaining Black Box Predictions in Medical AI

Brian Huynh
bth001@ucsd.edu

Joshua Lee
jdlee@ucsd.edu

Zoya Hasan
zohasan@ucsd.edu

Albert Hsiao
a3hsiao@health.ucsd.edu

Abstract

Pulmonary edema is a life-threatening condition in which fluid accumulates in the lungs, often requiring rapid diagnosis to prevent severe clinical outcomes. While deep learning models have demonstrated strong performance in medical imaging and clinical text analysis, their lack of interpretability limits trust and adoption in real clinical environments. In this work, we developed a multimodal framework that combines convolutional neural networks (CNNs) and a domain-specific large language model (LLM) to analyze chest radiographs and radiology reports while providing interpretable explanations for model predictions. For the imaging component, we train VGG16 and ResNet50 architectures to predict N-terminal pro-B-type natriuretic peptide (BNPP) biomarker values from chest X-ray images. To interpret model behavior, we apply Grad-CAM visualization and image ablation analysis to identify spatial regions that most strongly influence predictions. For the text component, we fine-tune the MediPhi medical language model using Low-Rank Adaptation (LoRA) to classify pulmonary edema severity from radiology reports. Model explanations are generated using LIME to identify influential linguistic features and cosine embedding analysis to examine semantic organization within the model's internal representation space. Our results show that CNN attention consistently focuses on clinically relevant regions of the chest, particularly around the cardiac and perihilar lung areas, while the language model relies on meaningful clinical descriptors associated with edema severity. These findings suggest that both imaging and language models learn patterns that align with established medical reasoning when paired with appropriate interpretability tools. Overall, this work demonstrates how multimodal explainability methods can help reduce the "black-box" nature of medical AI systems and improve transparency in AI-assisted clinical decision support.

Website: https://joshleh.github.io/dsc180b-project-website/
Code: https://github.com/brianthuynh10/dsc180-capstone

1 Introduction . . . 3
2 Methods . . . 4
3 Results . . . 8
4 Discussion . . . 14
5 Conclusion . . . 16
6 Contributions . . . 17

1 Introduction

Pulmonary edema is a life-threatening condition characterized by fluid accumulation in the lungs, leading to impaired gas exchange and acute respiratory distress. Early detection and severity assessment are critical, as delayed recognition can result in rapid clinical deterioration. Chest radiographs and accompanying radiology reports are routinely used to monitor pulmonary edema, yet their interpretation remains challenging due to subtle visual cues, subjective language, and variability across clinicians. These challenges motivate the use of machine learning systems that can assist clinicians by providing consistent, rapid, and interpretable assessments.

In earlier experiments, we investigated the feasibility of predicting pulmonary edema severity from chest radiographs using deep convolutional neural networks trained with objective serum biomarkers (BNPP) as continuous labels. Using a large UC San Diego institutional dataset, we demonstrated that CNN-based models can learn meaningful relationships between radiographic features and biomarker-derived severity, while highlighting the importance of model capacity and computational trade-offs. Although these results validated the potential of biomarker-supervised imaging models, they primarily focused on predictive performance and did not fully address how such models could be interpreted or deployed at scale in real clinical settings.

In this work, we extend our prior work by emphasizing interpretability, multimodal reasoning, and clinical applicability. In addition to image-based modeling, we incorporate unstructured radiology reports to enable language-based severity prediction and explanation. Specifically, we combine CNN-based imaging models with a fine-tuned medical language model and apply multiple explainability techniques—including Grad-CAM, image ablation, LIME, and cosine embedding similarity analysis—to better understand how these models detect pulmonary edema. We fine-tune a domain-specific large language model on a curated set of annotated reports and apply explainability techniques to expose the features driving model decisions. These approaches aim to bridge the gap between black-box prediction and clinically meaningful interpretation.

Finally, we explore large-scale clinical imputation by deploying the fine-tuned language model across tens of thousands of unlabeled reports from the UCSD dataset. This system generates interpretable predictions for edema presence and severity. This system is designed to support faster physician review, highlight potentially overlooked clinical indicators, and serve as a transparent decision-support tool rather than a replacement for expert judgment. Together, this work demonstrates how combining vision and language models with explainability techniques can enhance trust, scalability, and utility in medical AI systems for high-risk diagnostic tasks.

2 Methods

2.1 Overview

To support clinically interpretable pulmonary edema detection, we structured our methodology around two complementary axes of explainability: (1) visual explanation of imaging-based predictions using attention mechanisms, and (2) textual explanation of language-model predictions using multiple post-hoc interpretability techniques. This multimodal strategy enables both spatial localization in chest radiographs and semantic transparency in radiology reports.

2.2 Dataset

All medical data used in this study were collected by the University of California, San Diego (UCSD), with patient identity protected through anonymized identifier codes. The imaging dataset consisted of approximately 30,000 chest X-ray scans at a resolution of 1024×1024 pixels and were stored in HDF5 files. Each image was paired with the corresponding patient's log-transformed N-terminal pro-B natriuretic peptide (BNPP) value, age, and associated radiology report in CSV files. In addition, a curated subset of approximately 2,000 radiology reports contained structured diagnostic labels for multiple conditions, including pleural effusion, pneumonia, and pulmonary edema. For each condition, labels specified both presence and descriptive characteristics as annotated by radiologists. This labeled subset was used for supervised fine-tuning of the language model.

2.3 Data Preprocessing

Image Preprocessing. To accommodate computational constraints during CNN training, all chest X-ray images were downsampled from 1024×1024 to 256×256 resolution. Images were converted to grayscale and normalized to the range [0,1].

Standardization. Log-transformed BNPP values were standardized using z-score normalization computed from the training set mean and standard deviation. These statistics were subsequently applied to validation and test sets to prevent data leakage.

Text Preprocessing. For language model training, the labeled subset of 2,000 radiology reports was partitioned into training, validation, and test splits. Reports were tokenized into subword token IDs using the MediPhi tokenizer prior to transformer-based fine-tuning.

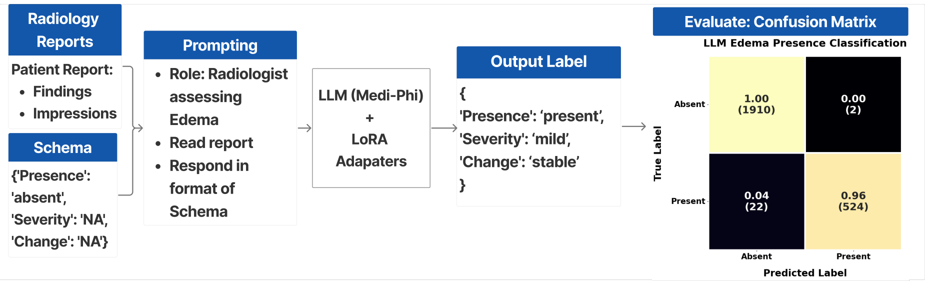

Radiology Reports
Patient Report:
• Findings
• Impressions
Schema
{'Presence': 'absent', 'Severity': 'NA', 'Change': 'NA'}
Prompting
• Role: Radiologist assessing Edema
• Read report
• Respond in format of Schema
LLM (Medi-Phi) + LoRA Adapaters
Output Label
{ 'Presence': 'present', 'Severity': 'mild', 'Change': 'stable' }
Evaluate: Confusion Matrix
LLM Edema Presence Classification
True Label
Absent
Present
1.00 (1910)
0.00 (2)
0.04 (22)
0.96 (524)
Absent
Present
Predicted Label

Figure 1: LLM (MediPhi) fine-tuning pipeline

2.4 Predictive Modeling

2.4.1 Language Model Fine-Tuning (MediPhi)

We fine-tuned a domain-specific large language model (MediPhi) using Low-Rank Adaptation (LoRA) on a labeled subset of radiology reports for multi-class pulmonary edema severity classification. Model performance was evaluated on a held-out test set using confusion matrices to assess agreement between predicted and ground-truth severity labels. After validation, the fine-tuned model was applied to the remaining unlabeled radiology reports to generate severity and edema predictions. These predicted labels were subsequently used to stratify the associated chest X-ray images for analysis across different sub-groups of edema patients.

2.4.2 CNN-Based BNPP Regression (ResNet50 and VGG16)

ResNet50 and VGG16 were adapted to perform regression by modifying their final layers to output a single continuous prediction corresponding to the standardized log-transformed BNPP value. Both models were trained using the Adam optimizer with an L1 loss function (mean absolute error). Hyperparameters included a batch size of 16, learning rate of 1×10^{-5}, and 50 training epochs. Model performance was evaluated using Pearson correlation (r) between predicted and ground-truth standardized BNPP values on the held-out test set.

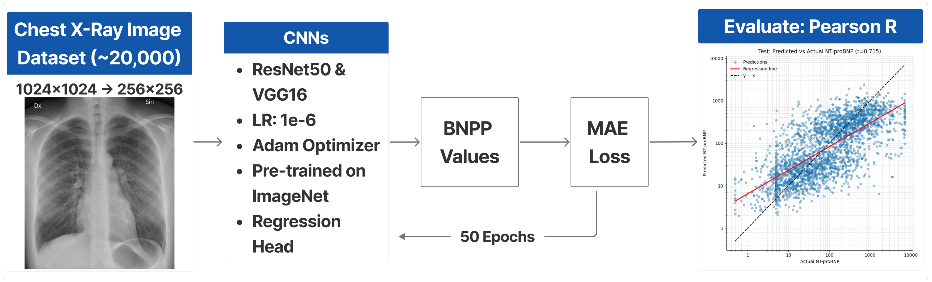


Figure 2: CNN (VGG16 & ResNet50 Training Process

2.5 LLM Explainability Methods

2.5.1 SHAP (SHapley Additive Explanations)

SHAP was used to interpret the model's binary probability output for absent versus present edema by assigning attribution scores to the tokenized input text. SHAP is based on Shapley values from cooperative game theory, where each input token is treated as a "player" contributing to the final prediction. For a given report, the method estimates how the model output changes when different subsets of tokens are included or removed, then attributes to each token its average marginal contribution across possible feature coalitions. In practice, this yields signed token-level importance values, where positive values indicate increased support for a class and negative values indicate reduced support. Because the model tokenizer decomposes text into subword units, SHAP explanations were computed at the subword level rather than strictly at the whole-word level, meaning a single clinical term could appear as multiple attributed pieces.

2.5.2 Cosine Similarity to Clinical Anchor Terms

To analyze the semantic structure learned by the language model beyond prediction-level attribution, we performed cosine embedding analysis on selected clinically meaningful terms and phrases. Cosine similarity measures the angular similarity between vector representations in embedding space, with higher similarity indicating that two terms are represented more closely by the model. We extracted embeddings for edema-relevant anchor phrases, such as descriptors of congestion, fluid burden, or explicit absence of edema, and then computed pairwise cosine similarities between these token representations. This produced a similarity matrix that allowed us to examine whether clinically related expressions clustered together and whether opposing concepts, such as disease presence versus negation, occupied distinct regions of latent space. Rather than explaining a single prediction, this analysis characterized how the fine-tuned model internally organized pulmonary edema terminology and whether that organization reflected clinically coherent semantic relationships.

2.5.3 LIME (Local Interpretable Model-Agnostic Explanations)

LIME was applied to the model's four-class severity output space to explain predictions across the categories absent, mild, moderate, and severe. Unlike SHAP, which computes additive contributions grounded in Shapley value theory, LIME builds a local surrogate model around an individual prediction. For each report, LIME generates perturbed versions of the original text by masking or removing subsets of tokens, queries the fine-tuned MediPhi model on these perturbed inputs, and observes how the predicted class probabilities change. It then fits an interpretable sparse linear model in the neighborhood of that example, using the perturbed samples and their corresponding outputs to approximate the model's local decision boundary. The coefficients of this surrogate model serve as token-level importance weights, indicating which words push the prediction toward or away from each severity class. Because LIME was run on the full four-class setting, it enabled class-specific interpretation of how different report terms influenced distinctions not only between edema absence and presence, but also between mild, moderate, and severe disease. Aggregating these local explanations across reports further allowed identification of broader global token importance patterns associated with each severity class.

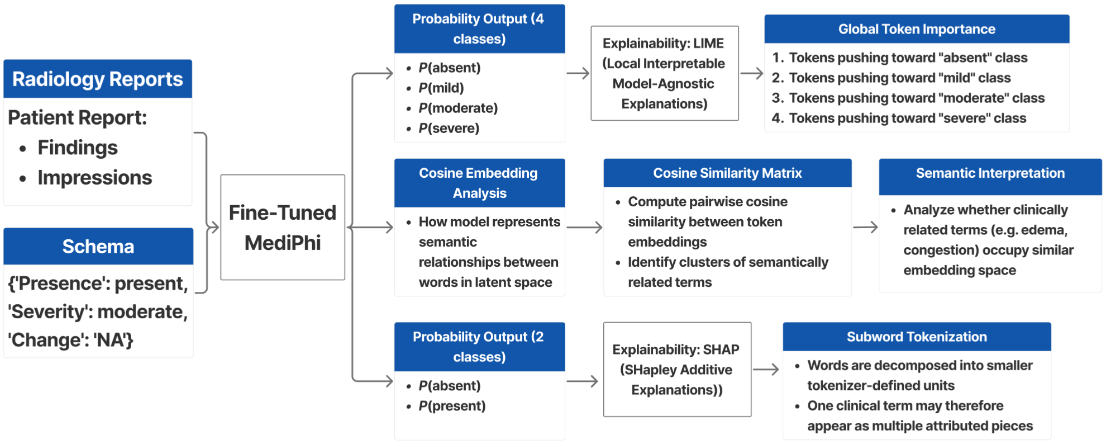

Radiology Reports
Patient Report:
• Findings
• Impressions
Schema
{'Presence': present, 'Severity': moderate, 'Change': 'NA'}
Fine-Tuned MediPhi
Probability Output (4 classes)
• P(absent)
• P(mild)
• P(moderate)
• P(severe)
Explainability: LIME (Local Interpretable Model-Agnostic Explanations)
Global Token Importance
1. Tokens pushing toward "absent" class
2. Tokens pushing toward "mild" class
3. Tokens pushing toward "moderate" class
4. Tokens pushing toward "severe" class
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
Semantic Interpretation
• Analyze whether clinically related terms (e.g. edema, congestion) occupy similar embedding space
Probability Output (2 classes)
• P(absent)
• P(present)
Explainability: SHAP (SHapley Additive Explanations))
Subword Tokenization
• Words are decomposed into smaller tokenizer-defined units
• One clinical term may therefore appear as multiple attributed pieces

Figure 3: LLM Explainability Workflow.

2.6 CNN Explainability Methods

2.6.1 Gradient-Weighted Class Activation Mapping (Grad-CAM)

We applied the Grad-CAM algorithm to both trained CNN architectures to visualize image regions contributing most strongly to BNPP predictions. For each CNN, Grad-CAM was computed with respect to the scalar regression output and applied to selected convolutional layers, including early, intermediate, and final layers to assess hierarchical feature localization. The resulting activation maps highlight spatial regions that most strongly influence the model's output. These maps were resized to match the input image resolution, normalized, and overlaid on the original chest X-ray for visualization. Grad-CAM analysis was performed across X-ray subgroups stratified by LLM-predicted severity labels (absent, mild, moderate, severe) to examine whether spatial attention patterns varied across subpopulations.

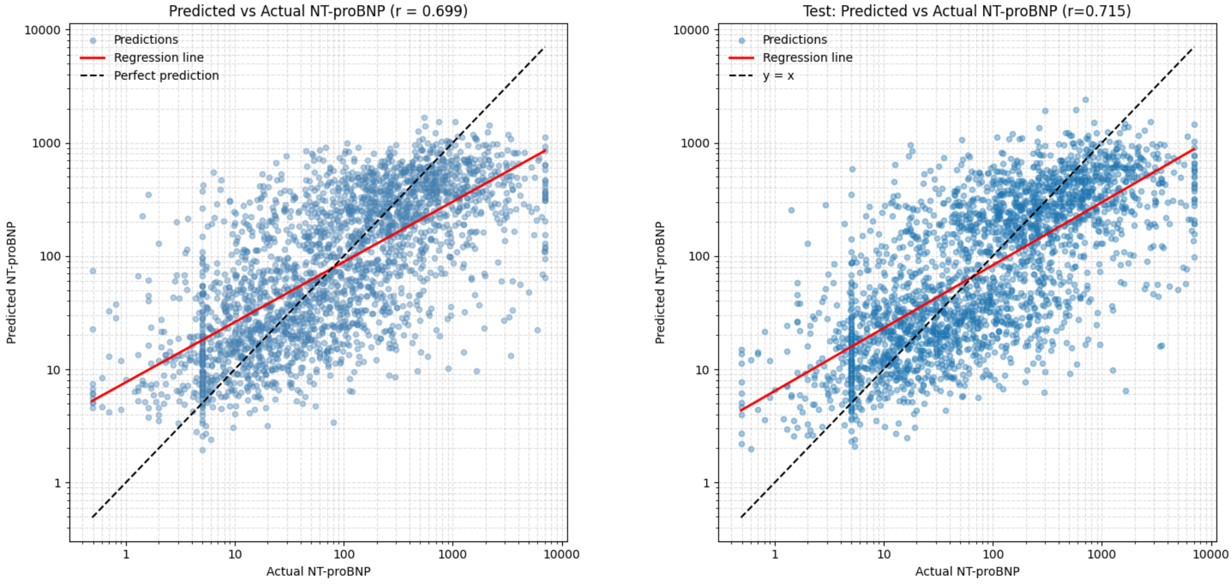
2.6.2 Image Ablation

We performed an image ablation to quantify model sensitivity to specific regions of the input image. For each trained CNN, a 16×16 occlusion patch was applied over the 256×256 input image with a stride of 16 pixels. At each spatial location, the selected patch was replaced with the mean pixel intensity of the image to approximate information removal while minimizing artificial edge artifacts or distribution shifts introduced by extreme pixel values. The ablated image is then passed through the CNN, and the resulting change in predicted

BNPP value was recorded. To evaluate global spatial importance, we measured the change in Pearson correlation (r) on the test set when ablating each grid region across all images. These performance drops were aggregated to generate spatial sensitivity heatmaps. To visualize these changes, ablation sensitivity maps were rendered as alpha-blended overlays on a representative chest X-ray image to highlight regions most strongly affecting model predictions. These perturbation-based maps were compared qualitatively with Grad-CAM visualizations to learn spatial regions influencing each model's predictions.

3 Results

3.1 CNN Results



(a) VGG16 predicted vs. true BNPP values

(b) ResNet50 predicted vs. true BNPP values

Figure 4: Regression performance of CNN models predicting BNPP values from chest X-rays.

3.1.1 Predictive Performance

Both CNN architectures demonstrated reasonable predictive performance on the held-out test set. As shown in Fig. 4, ResNet50 achieved a Pearson correlation of $r \approx 0.715$ between predicted and ground-truth standardized BNPP values, slightly outperforming VGG16, which achieved $r \approx 0.699$.

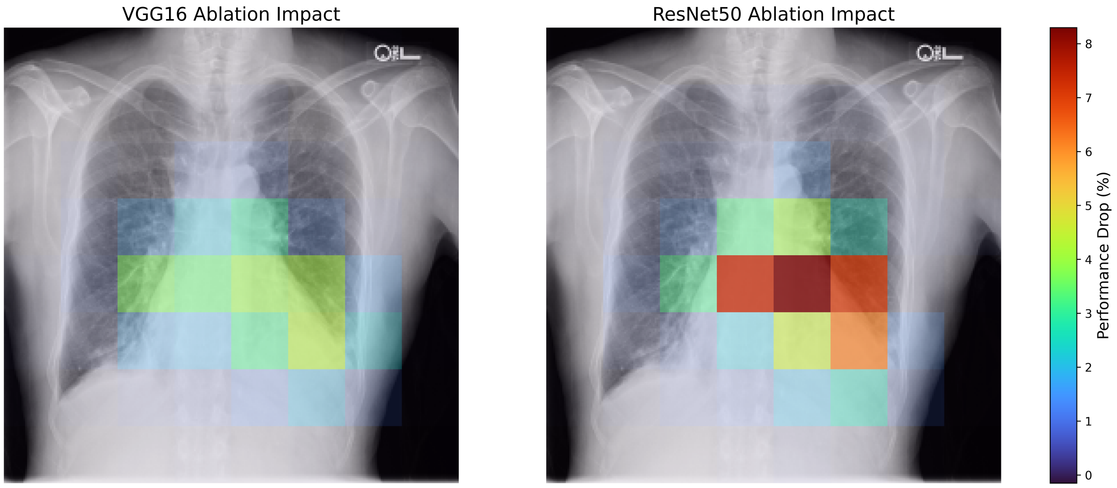


Figure 5: Comparison between VGG16 and ResNet50 performance drop when ablating image regions.

3.1.2 Image Ablation Sensitivity

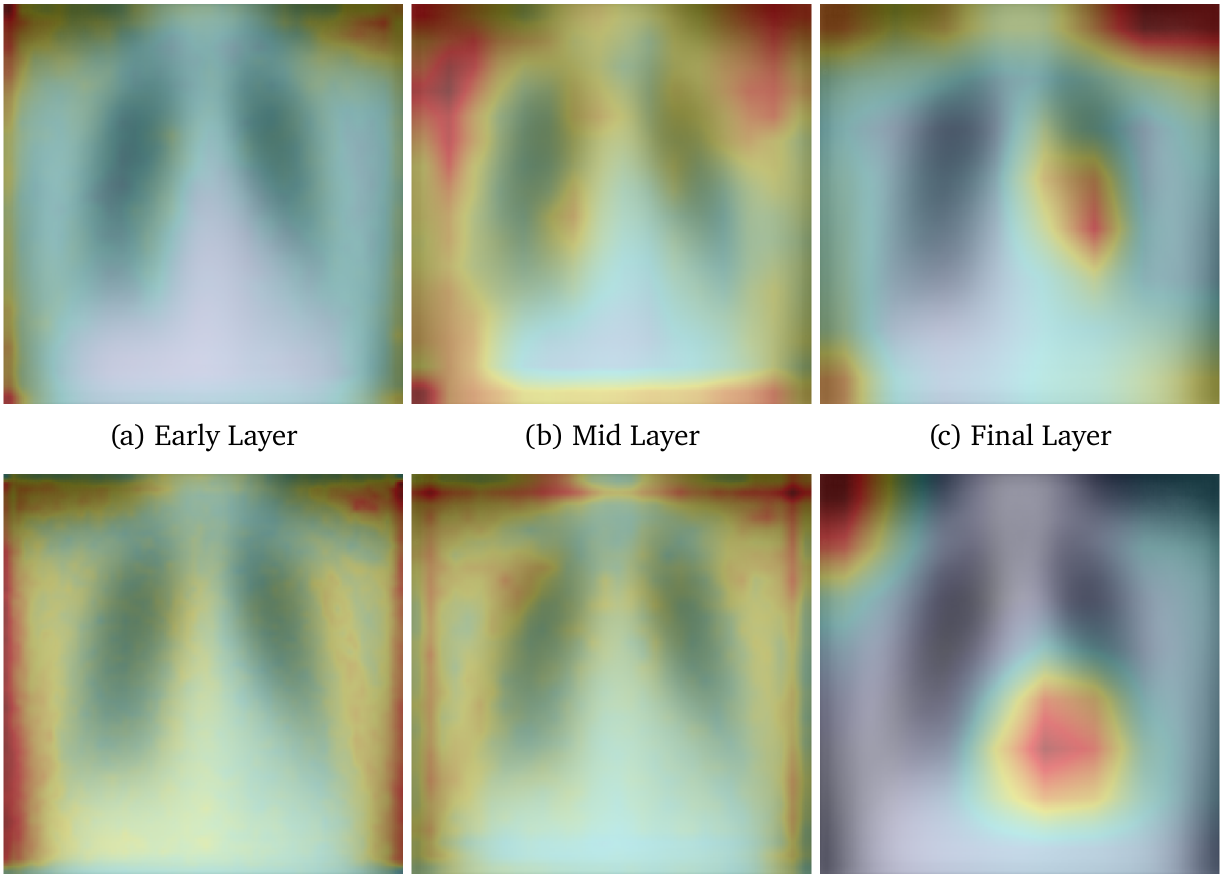
Image ablation analysis revealed performance degradation when occluding specific spatial regions. For both models, the largest decreases in Pearson correlation occurred when ablating around the cardiac and perihilar regions. ResNet50 exhibited a maximum performance drop of approximately 0.05 (from $r \approx 0.715$ to $r \approx 0.66$), while VGG16 showed a maximum decrease of approximately 0.036 ($r \approx 0.699$ to $r \approx 0.662$). Peripheral image regions produced comparatively little to no change in Pearson correlation.

3.1.3 Layer-wise Grad-CAM Analysis

Grad-CAM visualizations revealed architectural differences in spatial attention patterns. As shown in Fig. 6, VGG16 produced relatively sparse and sometimes scattered activation regions, including occasional attention outside the central chest region, with activations appearing near the corners of the X-ray images. In contrast, ResNet50 demonstrated more consistent activation patterns concentrated around the cardiac and left lung regions. Across both architectures, earlier convolutional layers exhibited diffuse attention distributed across the image. As depth increased, attention became more localized. In the final convolutional layers, both models showed greater concentration in anatomically relevant regions associated with pulmonary edema.

3.1.4 Severity-Stratified Attention Patterns

When stratified by LLM-predicted severity groups, both ResNet50 and VGG16 maintained relatively consistent attention patterns across mild, moderate, and severe cases, with activations concentrated around the upper lung fields and cardiac regions, seen in Fig. 7. In contrast, for cases labeled as absent edema, both models showed greater attention toward



(a) Early Layer

(b) Mid Layer

(c) Final Layer

(d) Early Layer

(e) Mid Layer

(f) Final Layer

Figure 6: Grad-CAM visualizations across network depth where CAM heatmaps were averaged then overlaid on averaged X-ray scans. Top row: VGG16 layers (early, mid, final). Bottom row: ResNet50 layers (early, mid, final). Red indicates stronger activations, while blue-green indicates weaker activations

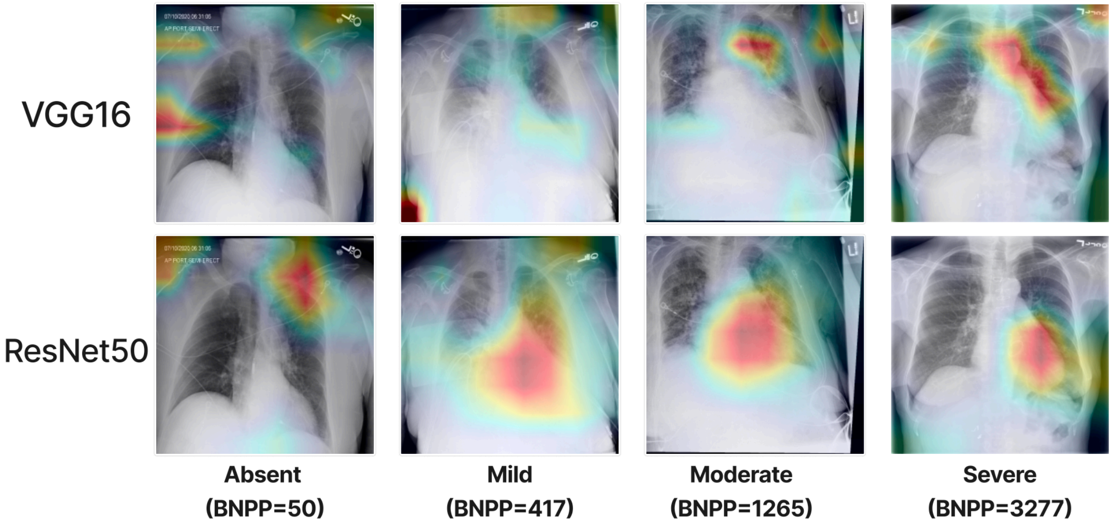


Figure 7: Grad-CAM attention maps for both models stratified by edema severity groups. Red indicates stronger activations, while blue-green indicates weaker activations.

outer regions of the chest rather than the central thoracic areas observed in present cases. Differences in spatial focus were also observed between architectures. VGG16 tended to produce broader, more widespread activation across the chest, whereas ResNet50 exhibited more localized attention focused on fewer regions.

3.2 LLM Explainability Results

3.2.1 SHAP Token Attribution

Initial explainability experiments using SHAP produced noisy and fragmented token attributions that were difficult to interpret at a clinical level. The most influential tokens identified by SHAP were often partial word fragments (e.g., subword tokens such as “ema”, “pul”, or “ed”), reflecting the underlying tokenizer rather than meaningful medical concepts. While some edema-related fragments appeared among the highest-weighted tokens for the “present” class, the explanations lacked coherent semantic structure and frequently highlighted formatting or report boilerplate. As a result, SHAP explanations provided limited clinical interpretability and motivated the use of alternative approaches better suited for text-based LLM analysis.

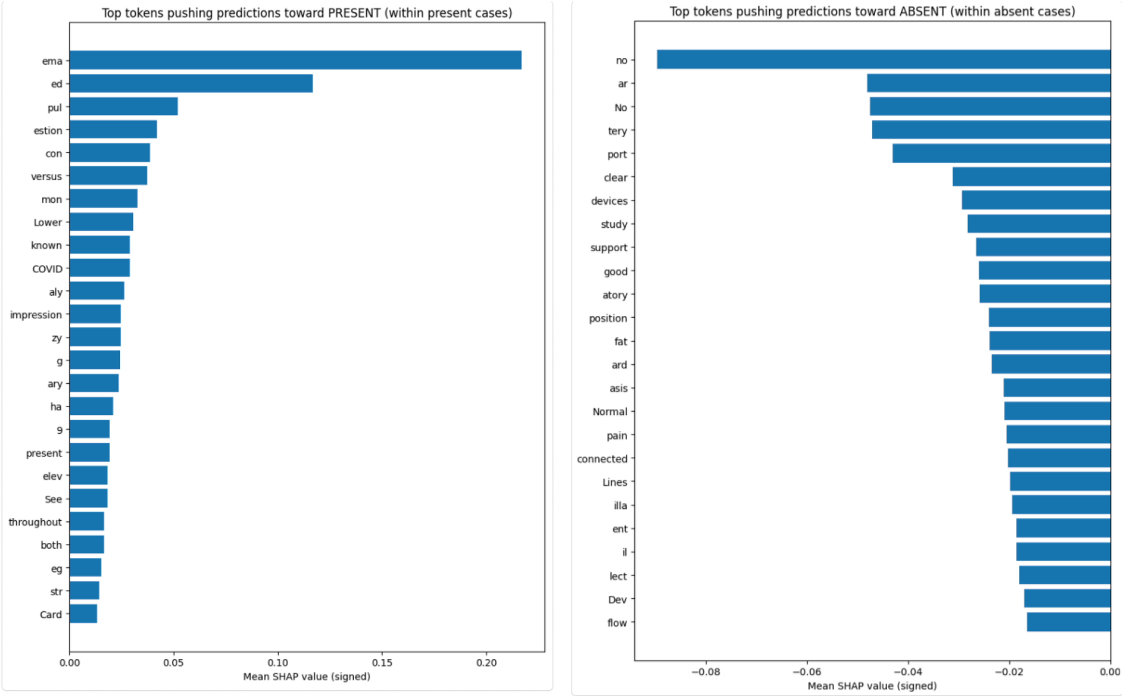


Figure 8: SHAP token attribution for pulmonary edema classification.

3.2.2 Cosine Embedding Analysis

To examine the model's internal semantic representation of clinically relevant terms, cosine similarity was computed between token embeddings extracted from the MediPhi model. The resulting similarity matrix reveals how the model organizes medical concepts in its embedding space. Certain terms form meaningful relationships, such as strong similarity between congestion and bone (0.76), suggesting co-occurrence patterns in radiology language, while strong negative similarity between congestion and pneumonia (-0.77) indicates the model differentiates between distinct pulmonary conditions. However, the expected semantic clustering between edema and congestion is relatively weak (-0.45), suggesting that the model's embedding space may reflect general linguistic structure rather than strongly specialized medical concept grouping. Overall, cosine similarity analysis provides insight into the model's internal representation of clinical terminology independent of prediction behavior.

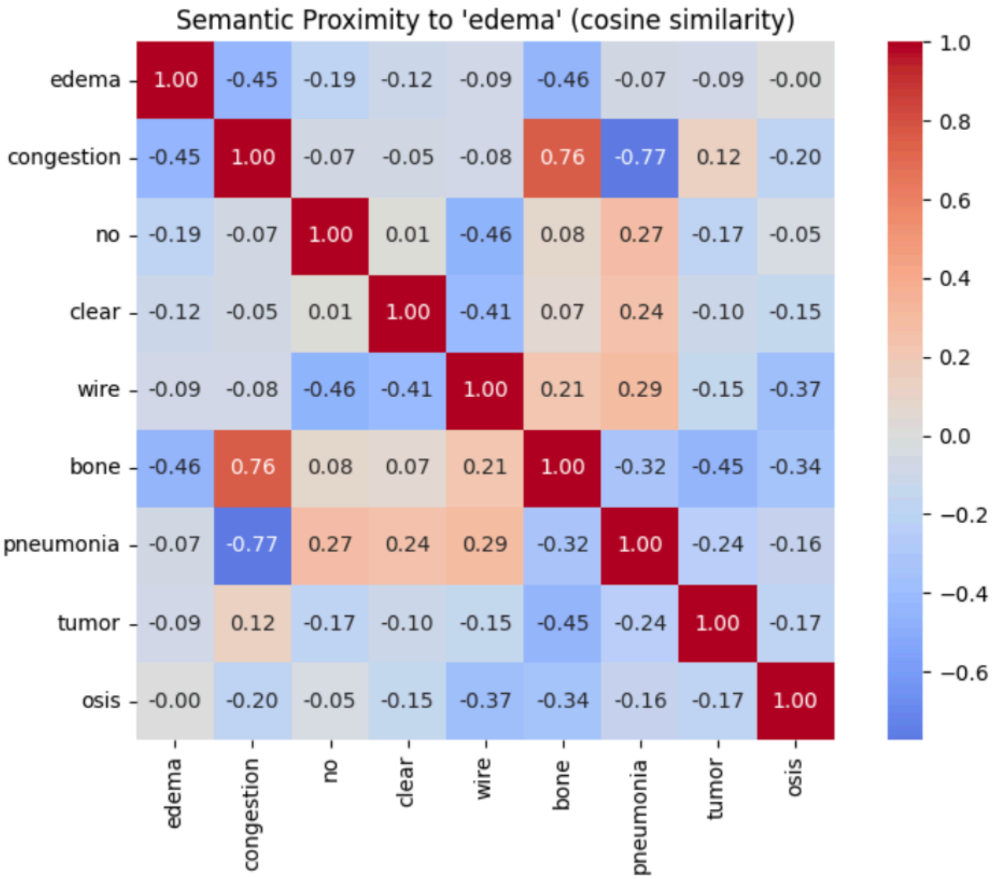


Figure 9: Cosine similarity between selected clinical terms in the model's embedding space.

3.2.3 LIME Global Token Attribution

To better understand which textual features influence model predictions, Local Interpretable Model-Agnostic Explanations (LIME) were applied across the test set and aggregated to produce global token importance scores for each edema severity class. The results show that clinically meaningful descriptors strongly influence predictions. Tokens such as edema, mild, moderately, and severe have the largest positive contributions toward their corresponding severity classes, indicating that the model relies heavily on explicit severity terminology present in radiology reports. Additionally, terms such as congestion, extensive, and worsening contribute to higher severity predictions, while words associated with absence or possibility shift predictions toward lower severity categories. Compared to SHAP, the aggregated LIME explanations produce more coherent and clinically interpretable patterns, demonstrating that the model primarily leverages explicit radiological descriptors and severity modifiers when classifying pulmonary edema severity.

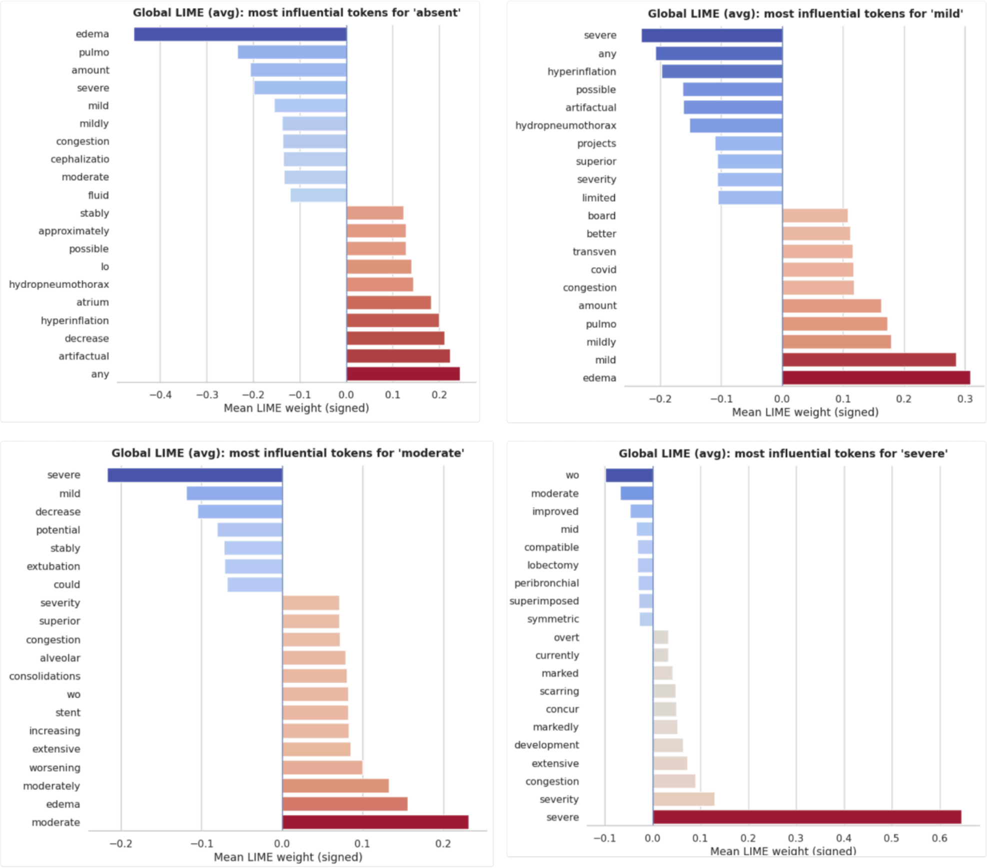


Figure 10: Aggregated LIME token contributions across the test set for each edema severity class.

4 Discussion

4.1 CNNs and Learned Spatial Regions

Our Grad-CAM and image ablation analyses suggest that both CNN architectures learned spatial patterns consistent with clinically relevant regions for pulmonary edema detection. Activation maps frequently highlighted the perihilar and cardiac regions rather than peripheral or background areas, indicating that the models were not relying on unrelated parts of the image. Grad-CAM visualizations also revealed architectural differences. VGG16 produced broader and more diffuse attention regions, while ResNet50 demonstrated more localized and anatomically focused activations. This difference may reflect variations in model depth and hierarchical feature learning.

Across LLM-predicted severity subgroups, both CNNs exhibited relatively stable activation patterns. ResNet50 maintained consistent localization across mild, moderate, and severe cases, with the absent case showing different attention patterns, possibly due to the lack of edema-related radiographic features. VGG16 showed similar activation regions for patients with edema (mild, moderate, and severe), with attention around the upper left lung and cardiac regions. In contrast, for absent cases, VGG16 displayed weaker activation near the heart and stronger attention outside the chest region, potentially for the same reason observed in the ResNet50 absent case.

Image ablation further supported these findings. Occluding regions around the heart and left lung resulted in the largest decreases in Pearson correlation, whereas peripheral regions produced minimal performance change. Together, these results suggest that the CNNs relied on spatial regions that are clinically meaningful for assessing pulmonary edema.

4.2 LLMs and Learned Words

This analysis of LLMs and learned words showed that language models do not simply rely on isolated keywords, but instead learn broader semantic and contextual relationships that shape their predictions. While token-level explainability methods such as LIME helped highlight which words most strongly influenced classification decisions, the learned word representations revealed that clinically related terms were also organized in meaningful ways within the embedding space. Together, these findings suggest that the model captures more than surface-level patterns, combining local word importance with deeper semantic structure. This is especially valuable in a clinical setting, where trust and interpretability are essential, because it shows that the model's reasoning is tied not only to specific report terms but also to medically relevant language relationships.

4.3 Limitations and Future Directions

This study has several limitations. First, the dataset was derived from a single institution and models were trained at a reduced image resolution (256×256), which may limit the ability of CNNs to capture fine-grained radiographic features. Training at higher resolutions could improve spatial specificity, as demonstrated in Huynh et al. (2022). Second, although MediPhi achieved strong performance on the labeled subset of radiology reports, when applying the model to the larger set of reports associated with our X-ray images there was no direct method to verify the certainty of the predicted labels. While BNPP values provide a related biomarker, there is no universally accepted threshold for binning BNPP values to edema severity categories. Future work could explore learning or calibrating these thresholds to better align imaging, biomarker, and language-derived labels. Third, while interpretability methods suggested alignment with clinically relevant spatial regions and terminology, these approaches remain largely qualitative. Future work should focus on developing quantitative measures to evaluate the alignment between model attention patterns and clinically annotated features. Finally, the imaging dataset lacked consistent normalization, with some images containing rotations or incomplete views of the chest. These inconsistencies may affect CNN training and influence the spatial features learned by the models.

5 Conclusion

This project explores how explainability techniques can improve transparency in medical AI systems designed to detect pulmonary edema from both imaging and clinical text. By combining convolutional neural networks for chest X-ray analysis with a fine-tuned medical language model for radiology report interpretation, we developed a multimodal framework capable of producing interpretable predictions across both visual and textual modalities.

Our results show that the CNN models learned spatial patterns consistent with clinically relevant regions associated with pulmonary edema. Grad-CAM visualizations and image ablation experiments consistently highlighted the cardiac and perihilar lung regions as the most influential areas for prediction, suggesting that the models rely on meaningful radiographic features rather than spurious image artifacts. Among the architectures tested, ResNet50 demonstrated more concentrated and anatomically consistent attention patterns compared to VGG16.

For the language model component, interpretability analyses revealed that the fine-tuned MediPhi model relies on clinically meaningful linguistic cues when predicting edema severity. LIME explanations highlighted descriptive terms related to congestion, progression, and severity as strong contributors to higher-severity predictions, while words associated with negation or stability contributed to lower-severity classifications. Embedding similarity analysis further demonstrated that the model organizes clinically related terminology into coherent semantic clusters, suggesting that its internal representation space captures meaningful medical relationships.

Taken together, these findings indicate that both imaging and language models learn patterns consistent with established clinical reasoning when paired with appropriate interpretability tools. By making these decision processes visible, our work helps address the "black-box" challenge in medical AI and highlights the importance of explainability for building trust in clinical decision-support systems.

Future work may extend this framework by incorporating larger multi-institution datasets, training models at higher image resolutions, and developing quantitative metrics to evaluate the alignment between model explanations and clinically annotated features. Ultimately, integrating interpretable multimodal AI systems into clinical workflows could support faster triage, improved diagnostic consistency, and more transparent AI-assisted healthcare.

References

Huynh, Justin, Samira Masoudi, Abraham Noorbakhsh, Amin Mahmoodi, Seth Kligerman, Andrew Yen, Kathleen Jacobs, Lewis Hahn, Kyle Hasenstab, Michael Pazzani, and Albert Hsiao. 2022. "Deep Learning Radiographic Assessment of Pulmonary Edema: Optimizing Clinical Performance, Training With Serum Biomarkers." *IEEE Access* 10: 48577–48588. [Link]

6 Contributions

- Brian Huynh: Responsible for training the models that were used in our interpretability methods this quarter. Fine-tuned MediPhi on our radiology reports and trained ResNet50 and VGG16 on the X-ray images. Created a pipeline for running the GradCAM explainability method on our trained CNNs and will be working on image ablation next. For the report, wrote out the methods section for GradCAM and image ablation.
- Zoya Hasan: Developed and evaluated multiple LLM explainability approaches, including SHAP, LIME, and cosine embedding similarity matrices by using fine-tuned MediPhi LLM prediction outputs. Tested MediPhi on pneumonia labels to better understand model architecture and behavior, and created class-specific visualizations, confusion matrices, bar charts, and poster-ready figures to communicate results.
- Joshua Lee: Designed and implemented the project website. Wrote the abstract, introduction, and conclusion sections of the written report. Conducted exploratory data analysis on the imaging and radiology report datasets to understand BNPP distributions and dataset structure. Assisted with system planning and design.