Utrecht University

Graduate School of Natural Sciences 2025

Explainable AI in a Radiology Scenario – INFOMXAI

Deep Learning for Pleural Effusion Detection & Similarity Based Retrieval in Chest X-rays

Nikita Aksjonov, Abhinav Atmuri, Bar Melinarskiy, and Konstantinos Zavantias

Model Overview

Using deep learning models in medical image analysis requires high accuracy, interpretability, and reliability. In this study, we extract feature embeddings from **DenseNet-169** a CNN model [1] to rank medical images based on Cosine Similarity to a given test image and apply Grad-CAM to visualize model decision making. The results demonstrate that Cosine Similarity ranks images effectively, with normalized DCG (nDCG) score of 0.79, indicating good alignment with the expected ranking. Finally, Grad-CAM heatmaps highlight clinically relevant regions contributing to trusting more in AI-assisted diagnosis.

Model Training

Data Preprocessing & Filtering

- We used frontal chest X-rays for consistency and replaced uncertain labels (-1) with true (1) for credibility.

Image Processing:

- Resized images to 224×224, converted to grayscale, and normalized pixel values for stable training.

Data Augmentation:

- Applied horizontal flipping, brightness/contrast adjustment, elastic transformations, Gaussian blur, rotation, affine transformations, and random cropping.

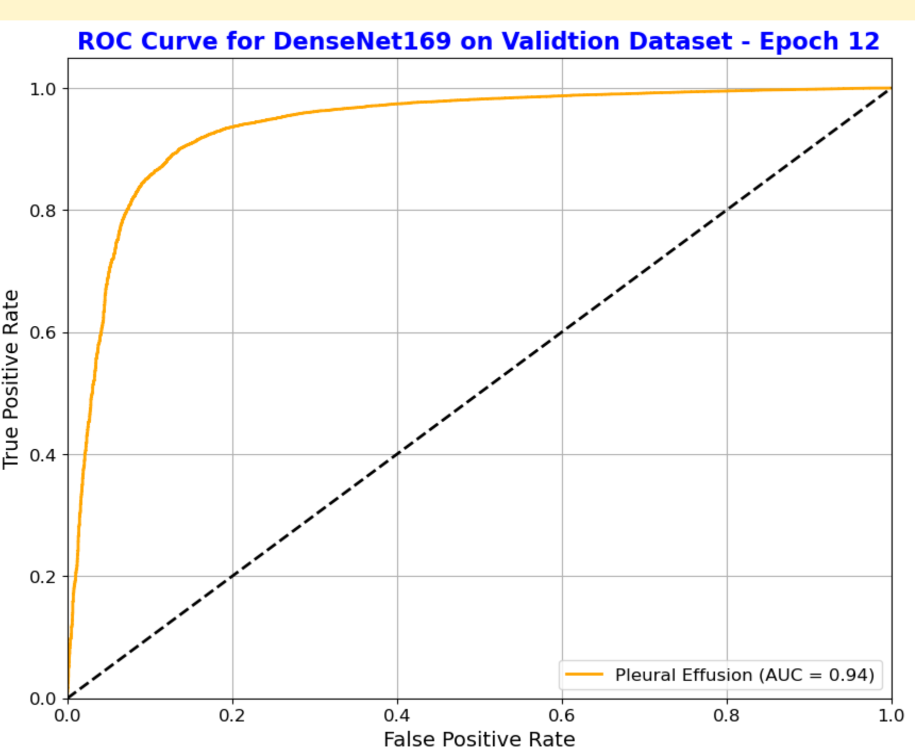
Training:

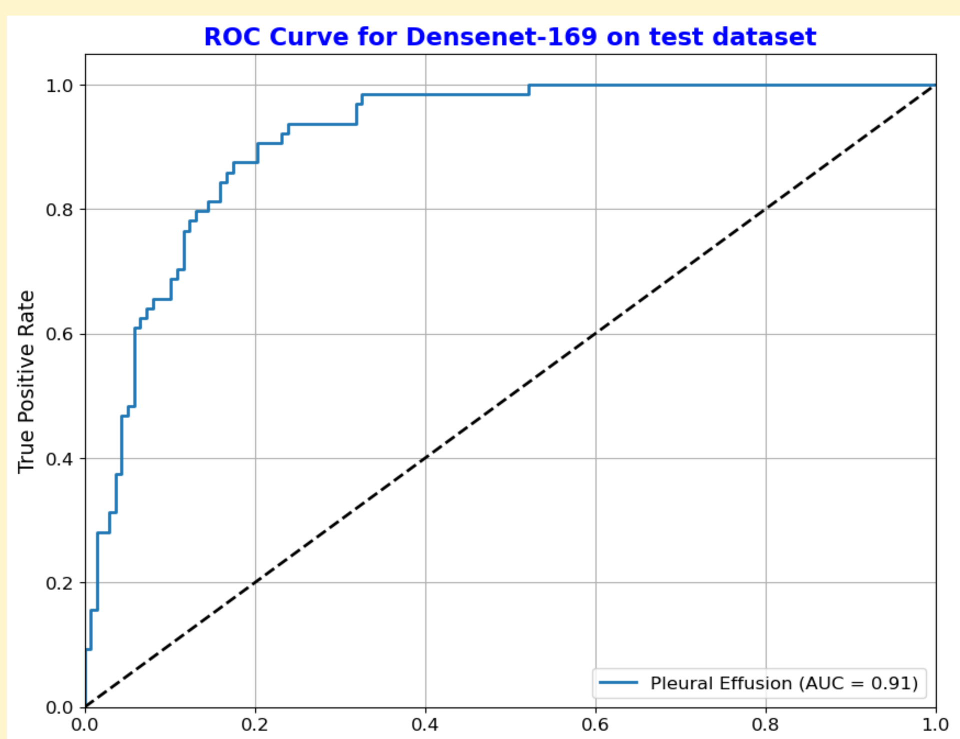
- **Loss Function:** BCEWithLogitsLoss (sigmoid + binary cross-entropy).
- **Optimizer:** AdamW (weight decay: 0.0001).
- **Learning Rate Scheduler:** ReduceLROnPlateau (factor: 0.1, patience: 5).
- **Epochs:** 25 (early stop at 12), Batch size: 32.
- **Performance Tracking:** Monitored loss curves, AUC-ROC, and accuracy; best model saved for final evaluation.

Training Results

High ROC indicates model's ability to distinguish negative and positive cases effectively after just the first epoch. Curve's shape shows that the model achieves high TPR while keeping low FPR, indicating learning meaningful patterns from early stage.
The table shows the model's performance improvement through epochs, as train loss decreased while accuracy increased. High AUC shows model's ability to maintain power even with the small loss.

Epoch	Train Loss	Train Accuracy	Test Loss	Test Accuracy	Test AUC
1	0.3393	0.8683	0.3212	0.8628	0.9417
12	0.2648	0.8991	0.3515	0.8713	0.9365





Explanations

Example-based

Retrieve and rank similar past cases to justify predictions, aligns with radiologists' diagnostic reasoning [2,3]. Inspired by content-based medical image retrieval [4] we use DenseNet169 to extract deep feature embeddings.

- Feature extraction for generation of embeddings for X-rays
- Cosine Similarity for comparison of new image against all trained
- Ranking of most similar images
- Experts' ratings manually ranked

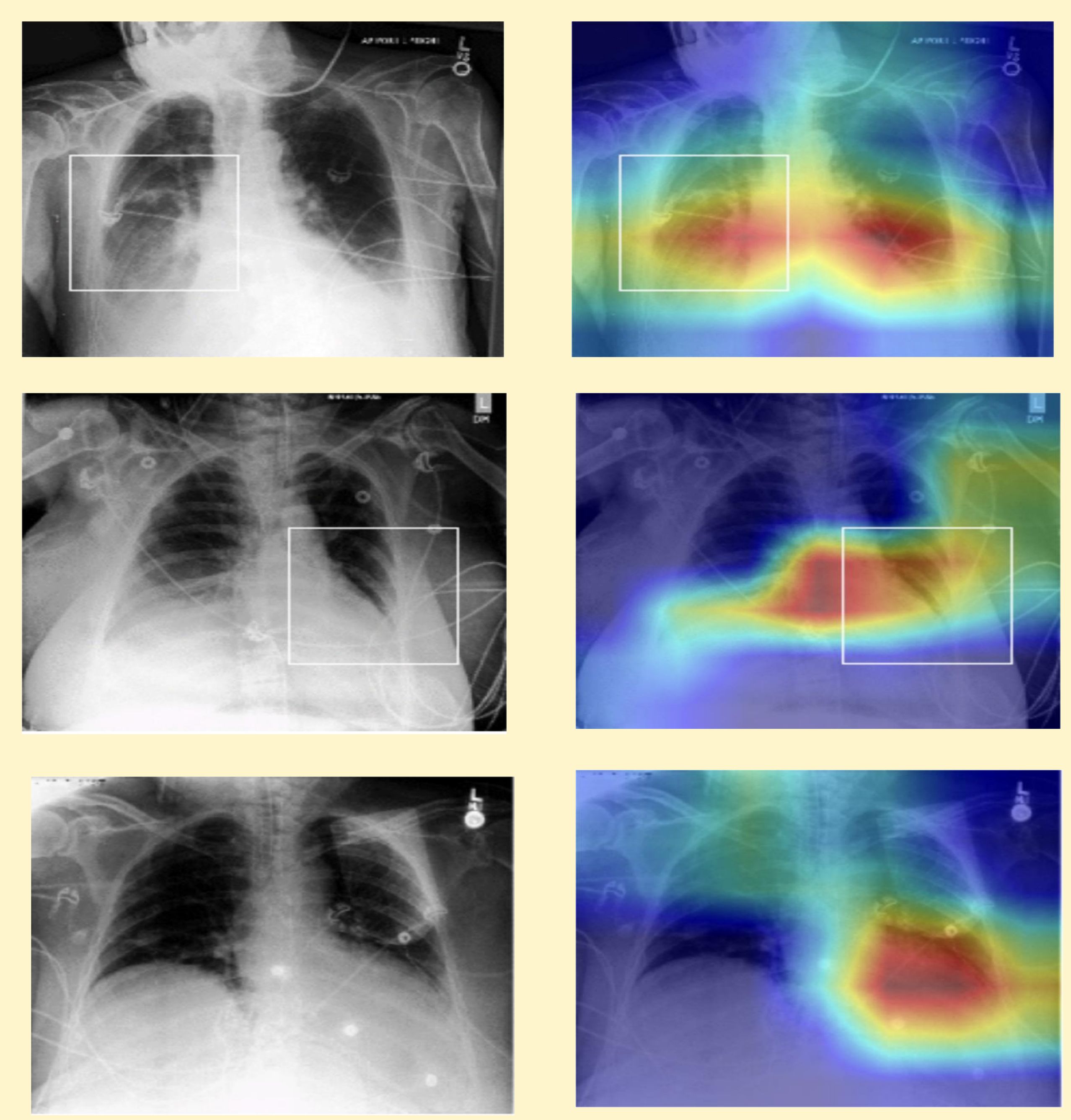
Saliency Map-based

Saliency maps highlight important regions influencing AI predictions and improving interpretability in medical imaging. In this project, **Grad-CAM** is used to generate heatmaps over X-rays and is validated against radiologist-annotated bounding boxes ensuring clinical relevance [5].

- **Grad-CAM** was applied to the last convolutional layer of **DenseNet-169** to generate class activation maps.
- Alignment with medically relevant regions by validation saliency maps against bounding box annotations.

Top 3 images most similar to the test image (Predictions - DenseNet169)

Original Image with Bbox | **Grad Cam Heatmap for Pleural Effusion**



Rad_3

Rad_2

Rad_4

Evaluate Explanations

Imaged	Similarity Score	Linear Ranking
Rad_3	0.91	5.5
Rad_2	0.89	5
Rad_4	0.88	4.5
Rad_1	0.86	4
Rad_6	0.82	3.5
Rad_5	0.79	3
Rad_8	0.67	2.5
Rad_7	0.32	2
Rad_10	0.25	1.5
Rad_9	0.23	1

nDCG Score - 0.79

Table 1

Represents the predicted ranking based on cosine similarity by extracting embeddings from DenseNet-169 in comparison to the test image. The obtained nDCG score of 0.79 indicates a fairly good ranking performance.

Metric	Quantus Metric Name	Avg Score	Description
Robustness	Local Lipschitz Estimate	2.05	The explanation is moderately sensitive to small input perturbations.
Faithfulness	Pixel Flipping	0.55	The explanation moderately aligns with the model's decision making.
Localisation	Relevance Mass Accuracy	0.18	This explanation does not greatly localize relevant features within the intended region of interest.
Complexity	Complexity	10.32	The explanation is relatively complex, which may reduce interpretability.

Table 2

The table summarises explanation quality metrics computed using Quantus, providing insights into how well the explanations align with model decisions and their interpretability.

Conclusion & Future Work

This study implements a pretrained DenseNet model for automated feature extraction and classification of X-Ray images. The model achieved high performance (AUC:0.94 for validation set, 0.91 for test set) and with Grad-CAM visualizations most relevant cases identified and the critical regions were highlighted.
This methodology offers a practical and scalable solution for medical image analysis, with potential application in computer-aided diagnosis and decision support systems.
Future work could focus on expanding datasets and integrating multimodal approaches.

References

1. Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
2. Rahman, M. D., et al. "Efficient Medical Image Retrieval Using DenseNet and FAISS for BIRADS Classification." arXiv preprint arXiv:2411.01473 (2024)
3. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*
4. Wang, Xiaosong & Peng, Yifan & Lu, Le & Lu, Zhiyong & Bagheri, Mohammadhadi & Summers, Ronald. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classificationand Localization of Common Thorax Diseases. arXiv:1705.02315. 10.48550/arXiv.1705.02315.

5. Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.