Explainable AI in a Radiology Scenario

Abhinav Atmuri *2556588* **Bar Melinarskiy** *2482975* **Konstantinos Zavantias** *7456123*
Nikita Aksjonov *7342195*

Abstract

Accurate and interpretable AI models are essential for medical diagnosis. This study presents an explainable deep learning model for Pleural Effusion detection in chest X-rays, integrating explainability techniques. Grad-CAM++ saliency maps were used to highlight important lung regions, and example-based retrieval ranks similar past cases to support decision-making. Radiologist-annotated bounding boxes used for validation of the results, indicating that the proposed approach maintains strong predictive performance while enhancing interpretability, aiding AI intergration in radiology.

1. Introduction

As deep learning models become increasingly common in the medical domain, a key challenge remains: despite their impressive accuracy, these models need to be interpretable to gain the trust of clinical professionals. This project addresses this need by developing an explainable deep-learning model for detecting pleural effusion in chest X-rays. DenseNet169 was selected as the backbone architecture because it provides an optimal balance between performance and explainability [7; 13].

To incorporate interpretability, two key approaches were implemented. First, an example-based retrieval system was developed that calculates cosine similarity using the model's final layer embeddings, allowing for the ranking of similar past cases when presented with a test image. These results were then visualized by highlighting critical lung regions using Grad-CAM++ saliency maps [10; 17]. Additionally, the model's explanations were validated by comparing them with ground truth bounding boxes annotated by radiologists, ensuring the approach aligned with clinical reasoning patterns [18].

2. Materials and methods

2.1. Dataset and Data Preprocessing

The CheXpert dataset [4; 14] was used to classify pleural effusion presence. The data was split into 80% training and 20% validation, with the training set containing 133,211 samples (97,815 positive, 35,396 negative, with uncertain cases treated as positive), resulting in class imbalance. A weighted sampler was implemented to address this by assigning higher selection probabilities to negative cases.

All images were resized to 224×224 pixels, converted to grayscale, and normalized. To improve generalization, the training set was augmented using Albumentations [9] with transformations such as horizontal flipping, elastic distortions, Gaussian blur, affine transformations, and random cropping. Finally, images were transformed into PyTorch tensors for model compatibility.

2.2. Model Training

The selected pre-trained DenseNet169 model [13] was trained with a batch size of 32 to efficiently process the dataset [7]. The model was initialized with pre-trained weights and additional custom layers were added to adapt it for detecting pleural effusion. No layers were frozen during training.

Training was performed using the BCEWithLogitsLoss function, suitable for binary classification, combined with a sigmoid activation for numerical stability [11]. The AdamW optimizer [15] (weight decay = 0.0001) was employed to mitigate overfitting, and the learning rate was adjusted using ReduceLROnPlateau, reducing it by a factor of 0.1 upon validation loss stagnation.

Although training was initially planned for 25 epochs, early stopping was applied at epoch 12 due to convergence. The model achieved a training accuracy of 86.83% at epoch 1, improving to 89.91% by epoch 12, while test accuracy increased from 86.28% to 87.13%. The test AUC slightly decreased from 0.9417 to 0.9365 between the first and final epochs, as shown in Fig. 1A.

Other architectures, including DenseNet201, were explored. While DenseNet201 achieved high accuracy, it exhibited poor saliency map performance, failing to highlight relevant regions in the X-rays.

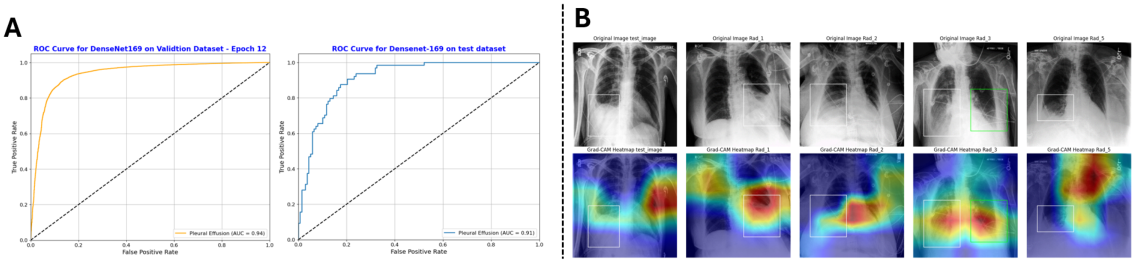

Figure 1: A. **ROC curves for DenseNet169** showing model performance on validation (AUC = 0.94, left) and test datasets (AUC = 0.91, right) for pleural effusion detection. B. **GradCAM++ saliency maps**. Top row: original resized images with marked bounding boxes (different colors indicate bilateral pleural effusion). Bottom row: corresponding GradCAM++ heatmaps.

2.3. Explanation methods

2.3.1. Saliency map-based evaluation

A Grad-CAM++ implementation was used to visualize the decision-making process of the model detecting pleural effusion in chest X-rays, improving feature localization by incorporating higher-order derivatives over Grad-CAM [17]. Ground truth images are converted to tensors, passed through the model to extract feature activations from the final convolutional layer, and then Grad-CAM++ computes a class-specific saliency map by backpropagating gradients to highlight key regions. The resulting visuals are normalized and aligned with the original image, showing the areas the model focuses on for classification.

2.3.2. Example based evaluation

To emulate the work of radiologists, cosine similarity was applied to feature embeddings extracted from the model's final layer to generate similarity scores between the test image and a set of ground truth-ranked images [10; 16]. Each ground truth image was compared to the test image, resulting in similarity scores that were subsequently sorted in descending order and linearly spaced for ranking.

3. Results and Discussion

To evaluate the model's classification performance, we incorporated radiologist-provided bounding box annotations into the images alongside heatmaps. These bounding boxes highlight the regions where pleural effusion is present, aiding in understanding the model's decision-making process. A sample image with bounding boxes and heatmaps is shown in Fig. 1B, demonstrating that the model, overall, accurately identifies pleural effusion regions, with notable exceptions such as the last image (Rad5) suggesting that more fine-tuning of the model is needed.

3.1. Example based evaluation

The results in Table 1 illustrate the similarity distribution among ranked images. Ranking quality was evaluated using the normalized Discounted Cumulative Gain (nDCG) score [8], yielding 0.795, which indicates approximately 80% alignment with radiologists' rankings. Although DenseNet201 achieved a superior nDCG score of 0.92, DenseNet169 was ultimately selected as the final model due to its superior performance in generating accurate saliency maps, which was prioritized for this application.

Metric	Rad_3	Rad_4	Rad_1	Rad_2	Rad_6	Rad_5	Rad_8	Rad_7	Rad_10	Rad_9
Similarity Score	0.928	0.915	0.897	0.877	0.86	0.804	0.664	0.377	0.283	0.25
Linear Ranking	5.5	5.0	4.5	4.0	3.5	3.0	2.5	2.0	1.5	1.0

Table 1: Similarity scores and linear rankings of images

3.2. Explanation Evaluation

The visual explanations generated by GradCAM++ were evaluated using Quantus metrics [12] to assess their interpretability and clinical relevance. The evaluation utilized input images (x_batch), prediction labels (y_batch), and grey-scale class activation maps (a_batch) to measure multiple aspects of explanation quality. The **robustness score** (0.699) measured via Local Lipschitz Estimation [3] indicates moderately robust explanations with acceptable sensitivity to input perturbations. **Faithfulness** (0.552), assessed through Pixel Flipping [5], suggests moderate alignment between explanations and the model's decision-making process. The low **localization** score (0.069) from Relevance Mass Accuracy [2] indicates that saliency maps inadequately focus on the most relevant image regions. The explanations exhibited high **complexity** [6] (9.044), potentially hindering interpretability, while the negative **randomization score** [1] (-0.025) suggests decreased explanation consistency when model parameters are randomized, raising reliability concerns.

4. Conclusion

This study implemented a pre-trained DenseNet model for pleural effusion detection in chest X-rays, achieving high classification performance. However, despite reasonable performance metrics, the generated GradCAM++ explanations demonstrated suboptimal localization scores and moderate faithfulness, indicating limitations in the model's interpretability. These findings suggest that while deep learning approaches can achieve strong diagnostic accuracy, further research is needed to enhance explanation quality through improved architectures, data preprocessing techniques, and augmentation strategies to develop clinically trustworthy AI systems that align with radiological ground truth.

References

[1] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*, 2018.

[2] Naveed Akhtar. A survey of explainable ai in deep visual modeling: Methods and metrics. *arXiv preprint arXiv:2301.13445*, 2023.

[3] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[4] Ashery. Chexpert dataset, 2023. URL https://www.kaggle.com/datasets/ashery/chexpert/data. Accessed: 2025-03-09.

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[6] Natalia C Berry, Laura Mauri, Philippe Gabriel Steg, Deepak L Bhatt, Stefan H Hohnloser, Matias Nordaby, Corinna Miede, Takeshi Kimura, Gregory YH Lip, Jonas Oldgren, et al. Effect of lesion complexity and clinical risk factors on the efficacy and safety of dabigatran dual therapy versus warfarin triple therapy in atrial fibrillation after percutaneous coronary intervention: a subgroup analysis from the redual pci trial. *Circulation: Cardiovascular Interventions*, 13(4):e008349, 2020.

[7] Keno K Bressem, Lisa C Adams, Christoph Erxleben, Bernd Hamm, Stefan M Niehues, and Janis L Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Scientific reports*, 10(1):13590, 2020.

[8] Róbert Busa-Fekete, György Szarvas, Tamás Elteto, and Balázs Kégl. An apple-to-apple comparison of learning-to-rank algorithms in terms of normalized discounted cumulative gain. In *ECAI 2012-20th European Conference on Artificial Intelligence: Preference Learning: Problems and Applications in AI Workshop*, volume 242. Ios Press, 2012.

[9] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alexandr Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *arXiv preprint arXiv:1809.06839*, 2020. URL https://arxiv.org/abs/1809.06839.

[10] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. page 1721–1730, 2015. doi: 10.1145/2783258.2788613. URL https://doi.org/10.1145/2783258.2788613.

[11] PyTorch Contributors. Bcewithlogitsloss, 2021. URL https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html. Accessed: 2025-03-09.

[12] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

[14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. URL https://arxiv.org/abs/1711.05101.

[16] D Sejal, T Ganeshsingh, KR Venugopal, SS Iyengar, and LM Patnaik. Image recommendation based on anova cosine similarity. *Procedia Computer Science*, 89:562–567, 2016.

[17] Shafiullah Soomro, Asim Niaz, and Kwang Nam Choi. Grad++scorecam: Enhancing visual explanations of deep convolutional networks using incremented gradient and score- weighted methods. *IEEE Access*, 12:61104–61112, 2024. doi: 10.1109/ACCESS.2024.3392853.

[18] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv:1705.02315*, 05 2017. doi: 10.48550/arXiv.1705.02315.