Multi-Paradigm Fusion of XAI Methods for Chest X-ray Explanations

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Deep learning models have shown strong performance in medical image analysis, particularly in chest X-ray interpretation, yet their adoption in clinical settings remains limited by confidence in model explanations. Although post-hoc explainability methods are widely used to inspect model behavior, their outputs are sensitive to methodological choices and can vary across model architectures and datasets. Gradient-based, perturbation-based, and game-theoretic methods capture distinct aspects of model behavior and exhibit different strengths and limitations, motivating their structured integration to obtain more stable attribution signals. In this paper, we propose a faithfulness-driven fusion framework that learns to combine gradient-based, perturbation-based, and game-theoretic explanations via a differentiable insertion–deletion objective, formalized through two complementary strategies: (i) hierarchical multi-scale pyramid fusion, and (ii) attention-guided fusion. We demonstrate the framework on ensemble architectures commonly used for chest X-ray analysis. Explanation quality is evaluated using insertion–deletion faithfulness tests and a controlled input perturbation analysis measuring explanation robustness. Experiments are conducted across two chest X-ray datasets: COVID-19 Radiography Database and NIH ChestXray14. Results demonstrate that the fusion strategies improve or remain competitive with single-method baselines in explanation faithfulness while maintaining comparable robustness under input perturbation, highlighting the value of combining complementary explanation paradigms for improving explanation stability in chest X-ray analysis.

Keywords: Explainable AI · Saliency Fusion · Chest X-ray Analysis

1 Introduction

Deep learning has fundamentally transformed medical image computing, achieving diagnostic accuracy that rivals expert performance across tasks ranging from disease detection to surgical planning. Despite these advances, clinical deployment remains limited by insufficient interpretability, as the largely black-box nature of deep models hinders trust in safety-critical medical settings [19]. Post-hoc explainable artificial intelligence (XAI) methods have emerged to attribute model predictions to specific input features [5], yet their outputs remain sensitive to methodological choices and vary substantially across architectures and

datasets. Prior studies report substantial variability in explanation fidelity across medical imaging applications, and demonstrate that saliency consistency can degrade markedly under realistic input perturbations [17, 19, 20]. When diagnostically relevant features are subtle and spatially diffuse such as ground-glass opacities or fine reticular patterns this sensitivity causes explanations to fluctuate across methods, undermining clinical utility [6, 1].

These challenges motivate a shift from selecting individual XAI methods to systematically integrating complementary signals. Grad-CAM [4] and Grad-CAM++ [9] (gradient-based), Integrated Gradients [13] (path-integrated attribution), LIME [12] and RISE [10] (perturbation-based and randomized masking), and SHAP [11] (game-theoretic) probe model behavior through distinct principles and exhibit different limitations [17, 26], motivating structured fusion to improve faithfulness and robustness over individual methods. These complementary characteristics motivate a unified framework for post-hoc XAI fusion. We investigate attention-guided and hierarchical multi-scale fusion strategies for post-hoc explainability.

We present a modular framework for multi-method XAI fusion through two strategies: (i) **attention-guided fusion**, which employs temperature-scaled softmax weighting to generate spatially adaptive combinations of explanation maps and (ii) **hierarchical multi-scale pyramid fusion**, inspired by pyramid architectures for dense prediction, which captures explanations at multiple resolutions to handle pathologies of varying spatial extent. We evaluate seven explanation methods (including proposed) spanning five paradigms on two chest X-ray datasets [14, 16]. Thus our contributions are threefold:

- A unified XAI fusion framework that addresses variability across attribution methods through two complementary strategies: attention-guided fusion and hierarchical multi-scale pyramid fusion.
- Two complementary parameterized fusion strategies that adaptively integrate attribution maps across methods and spatial scales using a differentiable faithfulness objective.
- Empirical evidence that fusion strategies with gradient-based weight optimization improve insertion-based faithfulness while maintaining competitive robustness, supporting structured multi-paradigm integration for more stable interpretability in chest X-ray analysis.

2 Related Work

2.1 Post-hoc XAI Methods and Reliability

Gradient-based methods (Grad-CAM [4], Grad-CAM++ [9], Integrated Gradients [13]), perturbation-based methods (LIME [12], RISE [10]), and the game-theoretic SHAP [11] are widely used in chest X-ray analysis [8, 6, 2, 30], yet these paradigms exhibit complementary failure modes [17, 20]: (i) gradient-based methods capture fine-grained texture cues but may suffer from gradient saturation; (ii) perturbation-based approaches model broader context but incur high

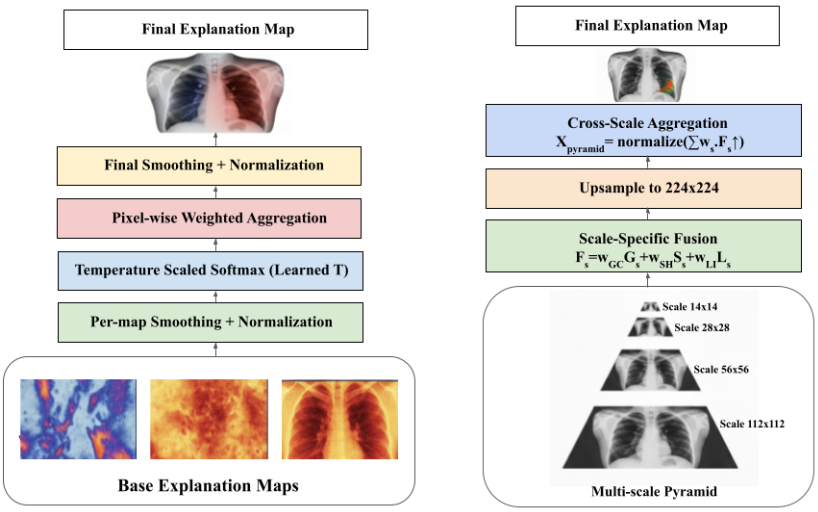

Fig. 1. Overview of the proposed multi-paradigm fusion framework. **Left:** Temperature-scaled softmax attention integrates multi-paradigm explanation maps to produce a spatially adaptive attribution map. **Right:** Hierarchical multi-scale pyramid fusion performs scale-specific integration followed by cross-scale weighted aggregation to capture both coarse and fine-grained pathological patterns.

computational cost at fine resolution; (iii) the game-theoretic method provides relatively stable global attributions yet struggles with precise localization of diffuse findings. Prior chest X-ray studies primarily evaluate explanation methods in isolation [7, 19], despite well-documented variability across attribution paradigms [17, 20, 27]. Recent work has explored saliency map fusion to improve weakly supervised semantic segmentation [3], while we investigate the fusion of post-hoc attribution methods for image classification.

2.2 Multi-scale Processing in Medical Imaging

Hierarchical architectures such as U-Net [28] and Pyramid Vision Transformers [29] demonstrate that multi-resolution representations are effective in medical image analysis. In chest radiography, pathological findings span multiple spatial scales [16], yet existing XAI approaches typically generate explanations at a single resolution [4, 20], overlooking hierarchical radiographic structure and potentially missing fine-scale diagnostic cues.

3 Methodology

3.1 Overview

Given an input radiograph $\mathbf{x} \in \mathbb{R}^{224\times224\times3}$ and a trained model f, we generate post-hoc explanations using Grad-CAM [4], Grad-CAM++ [9], Integrated Gradients [13], RISE [10], LIME [12], and SHAP [11], representing gradient-based,

Architecture	Type	Acc	F1	AUROC
COVID-19 Radiography				
Ensemble	-	**96.51**	**96.39**	**99.49**
DenseNet-201 [21]	CNN	89.57	90.50	98.48
ResNet-152 [22]	CNN	85.26	85.28	97.35
Swin [23]	Transformer	94.77	94.58	98.21
ConvNeXt [24]	CNN	93.73	92.87	97.98
NIH ChestXRay14				
Ensemble	-	**88.35**	**53.92**	**83.71**
DenseNet-201 [21]	CNN	81.28	30.98	79.18
ResNet-152 [22]	CNN	80.87	30.02	80.57
Swin [23]	Transformer	84.13	53.15	83.38
ConvNeXt [24]	CNN	84.46	52.87	83.59

Table 1. Architecture comparison across datasets. Best values per dataset are shown in bold. The ensemble model is selected for downstream XAI fusion experiments.

perturbation-based, and game-theoretic attribution paradigms. The resulting attribution maps are aligned and normalized prior to fusion. We investigate two fusion strategies: attention-guided fusion and hierarchical multi-scale pyramid fusion. Fusion parameters are optimized using a differentiable faithfulness objective. The fused explanation $\hat{\mathbf{X}}$ is evaluated using insertion–deletion faithfulness and robustness under input perturbations. For all the experiments, we use an ensemble of ViT-B/16 [25], Swin-B [23], and ConvNeXt-B [24], selected based on validation performance across the evaluated datasets (Table 1).

3.2 Attention-Guided Fusion

We employ temperature-scaled softmax attention for spatially adaptive fusion. Each explanation map $\mathbf{X}_i$ is Gaussian-smoothed,

$$\tilde{\mathbf{X}}_i = \mathcal{G}_\sigma[\mathbf{X}_i], \tag{1}$$

where $\mathcal{G}_\sigma$ denotes a 3×3 Gaussian filter. Attention weights at spatial location $\mathbf{p} = (h, w)$ are

$$\alpha_i(\mathbf{p}) = \frac{\exp\left(\tilde{\mathbf{X}}_i(\mathbf{p})/T\right)}{\sum_{j \in \mathcal{M}} \exp\left(\tilde{\mathbf{X}}_j(\mathbf{p})/T\right)}, \tag{2}$$

where $\mathcal{M}$ is the set of explanation methods and T controls the sharpness of the attention distribution. The temperature is optimized using the faithfulness objective described in Section 3.4. Intuitively, Eq. (2) assigns spatially varying weights to attribution methods through a temperature-scaled softmax, giving greater emphasis to stronger local attribution responses. Consequently, the fused explanation adapts to the relative contribution of each method at every spatial location rather than treating all methods equally. The fused explanation is

$$\mathbf{X}_{\text{fusion}}(\mathbf{p}) = \sum_{i \in \mathcal{M}} \alpha_i(\mathbf{p})\, \mathbf{X}_i(\mathbf{p}), \tag{3}$$

followed by Gaussian smoothing and normalization,

$$\hat{\mathbf{X}}_{\text{attention}} = \text{normalize}(\mathcal{G}_{\sigma'}[\mathbf{X}_{\text{fusion}}]) \,, \tag{4}$$

where $\sigma' = 0.3$.

3.3 Hierarchical Multi-Scale Pyramid Fusion

To capture findings at different spatial scales, we employ a hierarchical multi-scale fusion strategy over resolutions $\mathcal{S} = \{14, 28, 56, 112\}$. Grad-CAM, SHAP, and LIME are selected for pyramid fusion as they produce spatial attribution maps that are well suited to multi-scale resizing. In contrast, IG and RISE are retained as individual baselines because their pixel-level and stochastic attribution characteristics are less suited to repeated coarse-scale resampling. At each scale, $\mathbf{X}_i$ is resized to resolution s via a resize/resampling operator $\mathcal{R}(\cdot, s)$,

$$\mathbf{X}_{i,s} = \mathcal{R}(\mathbf{X}_i, s), \tag{5}$$

Grad-CAM [4], SHAP [11], and LIME [12] maps are fused using learned method weights

$$\mathbf{F}_s = w_{\text{GC}}\mathbf{G}_s + w_{\text{SH}}\mathbf{S}_s + w_{\text{LI}}\mathbf{L}_s, \tag{6}$$

where $\mathbf{G}_s$, $\mathbf{S}_s$, and $\mathbf{L}_s$ denote scale-specific explanation maps.

Each fused map is normalized and upsampled,

$$\mathbf{F}_s^{\uparrow} = \mathcal{R}(\text{normalize}(\mathbf{F}_s), 224) \,. \tag{7}$$

The final explanation is obtained by aggregating the upsampled maps using learned scale weights,

$$\hat{\mathbf{X}}_{\text{pyramid}} = \text{normalize}\left(\sum_{s\in\mathcal{S}} w_s \, \mathbf{F}_s^{\uparrow}\right) . \tag{8}$$

3.4 Learnable Fusion Weight Optimization

Fusion parameters are optimized using a differentiable faithfulness objective:

$$\mathcal{F}(\hat{\mathbf{X}}) = \text{AUC}_{\text{ins}}(\hat{\mathbf{X}}) + 0.5\big(1 - \text{AUC}_{\text{del}}(\hat{\mathbf{X}})\big). \tag{9}$$

Since hard top-k masking is non-differentiable, we replace it with a sigmoid soft mask,

$$m_t(\mathbf{p}) = \sigma\Big(\gamma\big(\hat{\mathbf{X}}(\mathbf{p}) - \tau_t\big)\Big) \,, \tag{10}$$

which approaches hard masking as $\gamma \to \infty$. The mask is applied to blurred/original image blends to construct insertion–deletion curves using 15 steps and trapezoidal integration. Optimization uses soft masking, whereas final evaluation uses hard masking (100 steps, zero baseline).

Method	Insert ↑	Delete ↓	SSIM ↑	RankCorr ↑
Gradient-based				
Grad-CAM [4]	0.630±0.158	0.498±0.231	**0.581±0.231**	0.269±0.265
Grad-CAM++ [9]	0.588±0.199	0.488±0.224	0.545±0.369	0.155±0.252
Integrated Gradients [13]	0.667±0.181	**0.419±0.266**	0.495±0.097	0.119±0.107
Perturbation-based				
RISE [10]	0.603±0.153	0.420±0.217	0.352±0.047	0.073±0.066
LIME [12]	0.664±0.166	0.474±0.187	0.323±0.066	0.332±0.147
Attention Fusion (Ours)	**0.670±0.144**	0.507±0.182	0.454±0.102	**0.449±0.218**
Pyramid Fusion (Ours)	**0.690±0.122**	0.524±0.192	0.563±0.129	**0.492±0.246**

Table 2. Explanation faithfulness and robustness on the COVID-19 Radiography dataset [14, 15]. Best values per column are highlighted in bold.

For pyramid fusion, method weights $\mathbf{w}^m$ and scale weights $\mathbf{w}^s$ are parameterized through sigmoid-normalization and optimized with

$$\mathcal{L} = -\mathcal{F}(\hat{\mathbf{X}}) + \lambda_{\text{TV}}\mathcal{L}_{\text{TV}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}, \tag{11}$$

combining smoothness, sparsity, scale-diversity, and method-entropy regularization.

Optimization uses AdamW (Learning Rate 1×10^{-2}, weight decay 1×10^{-4}), with the learning rate halved after three stagnant epochs and a maximum of 20 epochs.

3.5 Implementation Details

We evaluate on COVID-19 Radiography [14, 15] (21,165 images, 4 classes) and NIH ChestXray14 [16] (112,120 images, 15 pathology labels), using the official splits. Fusion parameters are optimized on the training split; all faithfulness and robustness metrics are computed on the held-out test set. Insertion–deletion AUC is computed over 100 steps using a zero baseline. Robustness is evaluated under Gaussian noise ($\epsilon \in \{0.01, 0.05, 0.10\}$), small rotations, center cropping, and contrast adjustments [17]; for NIH ChestXray14, explanations are generated for the highest-confidence predicted pathology. All experiments used an NVIDIA A100 GPU (40 GB). Models were trained for 50 epochs using AdamW (weight decay 1×10^{-5}, batch size 32), with learning rates of 3×10^{-4} for NIH ChestXray14 and 1×10^{-4} for COVID-19 Radiography; images were resized to 224×224. Integrated Gradients used 64 steps, RISE 6,000 masks, SHAP 100 background samples, and LIME 300 perturbations. All results used a fixed random seed of 42. For the COVID-19 Radiography dataset, pyramid fusion converged to method weights $\mathbf{w}^m = [0.396, 0.225, 0.379]$ for Grad-CAM, SHAP, and LIME, and scale weights $\mathbf{w}^s = [0.124, 0.193, 0.322, 0.361]$ for resolutions $\{14, 28, 56, 112\}$.

Method	Insert ↑	Delete ↓	SSIM ↑	RankCorr ↑
Gradient-based				
Grad-CAM [4]	0.585±0.146	0.611±0.163	0.418±0.184	0.276±0.234
Grad-CAM++ [9]	0.575±0.150	0.629±0.184	0.379±0.248	0.084±0.141
Integrated Gradients [13]	0.637±0.123	0.411±0.120	0.465±0.109	0.231±0.124
Perturbation-based				
RISE [10]	0.450±0.090	**0.309±0.073**	0.328±0.025	0.056±0.028
LIME [12]	0.508±0.142	0.435±0.115	0.357±0.063	0.285±0.127
Attention Fusion (Ours)	0.562±0.132	0.375±0.095	0.375±0.091	**0.286±0.182**
Pyramid Fusion (Ours)	**0.641±0.119**	0.408±0.106	**0.468±0.098**	**0.317±0.221**

Table 3. Explanation faithfulness and robustness on the NIH ChestXray14 dataset [16]. Best values per column are highlighted in bold.

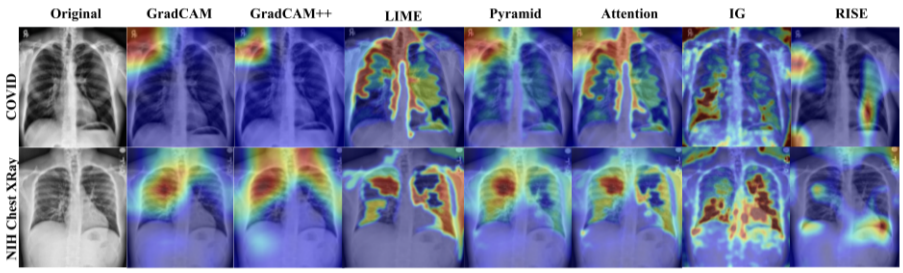

Fig. 2. Qualitative saliency maps on COVID-19 (top) and NIH ChestXray14 (bottom). Gradient-based methods yield highly localized peaks; perturbation-based maps are fragmented; both fusion strategies produce more spatially coherent distributions. Pyramid Fusion produces more spatially coherent attribution maps, consistent with its higher insertion AUC.

4 Results and Discussion

4.1 Explanation Faithfulness and Robustness

Tables 2 and 3 summarize insertion AUC, deletion AUC, and robustness metrics across both datasets. On COVID-19 Radiography [14, 15], Pyramid Fusion achieves the highest insertion AUC (0.690) and rank correlation (0.492), followed by Attention Fusion and Integrated Gradients [13]. On NIH ChestXray14 [16], Pyramid Fusion similarly leads on insertion AUC (0.641) and SSIM robustness (0.468), surpassing Integrated Gradients on both metrics. RISE retains the best deletion AUC on NIH (0.309), consistent with its masking-based attribution mechanism, while fusion strategies yield higher deletion AUC, consistent with their broader spatial coverage and the inherent trade-off between insertion faithfulness and deletion precision. Figure 2 illustrates qualitative differences between individual and fused explanations. Gradient-based methods produce localized responses, whereas perturbation-based approaches generate broader or more fragmented regions. Both fusion strategies produce more spatially coherent attribution maps.

Configuration	Insert AUC	Δ
Attention Fusion (G+S+L)	**0.670**	–
w/o SHAP [11]	0.582	−10.2%
w/o LIME [12]	0.593	−17.8%
w/o Grad-CAM [4]	0.525	−19.9%
Pyramid Fusion (learned)	**0.690**	–
Fixed init. $[0.396, 0.225, 0.379]$	0.671	−2.8%
Uniform weights	0.609	−11.7%

Table 4. Ablation study on the COVID-19 Radiography dataset [14] using insertion AUC. Δ indicates relative performance drop compared to the full configuration.

4.2 Ablation Study

Table 4 analyzes component contributions on COVID-19 Radiography [14]. Removing individual attribution methods consistently reduces performance, indicating complementary contributions. Excluding Grad-CAM [4] produces the largest decrease in insertion AUC (19.9%), followed by LIME [12] (17.8%) and SHAP [11] (10.2%). For Pyramid Fusion, learned weights achieve an insertion AUC of 0.690, outperforming both fixed initialization (−2.8%) and uniform weighting (−11.7%), demonstrating the benefit of faithfulness-driven optimization.

4.3 Discussion

The results indicate that attribution paradigms provide complementary information: gradient-based methods capture localized discriminative regions, while perturbation-based methods identify broader contextual evidence. Integrating these signals produces more coherent maps while preserving faithfulness. The consistent advantage of Pyramid Fusion across both datasets suggests multi-scale aggregation is effective for combining heterogeneous explanations, with learned scale weights favoring finer resolutions (scales 56 and 112), suggesting that mid-to-fine-scale information contributes more strongly within the proposed fusion framework. The ablation study confirms that each attribution source contributes non-redundant information and that learned weights outperform fixed or uniform alternatives, supporting structured multi-paradigm fusion as a practical approach to improving explanation stability without modifying the underlying model.

5 Limitations and Future Work

We present a modular framework for fusing complementary post-hoc attribution methods that improves explanation faithfulness while maintaining competitive robustness. Evaluation is limited to quantitative faithfulness and robustness metrics without clinical validation. Future work will investigate clinical validation through radiologist studies, extend the framework to CT and MRI, and explore lightweight fusion variants for clinical deployment.

References

1. Liu, B., Zou, K., Zhan, L., Lu, Z., Dong, X., Chen, Y., Xie, C., Cao, J., Wu, X.-M., Fu, H.: GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis. arXiv preprint arXiv:2411.16778 (2024)
2. Demir, U., Jha, D., Zhang, Z., Keles, E., Allen, B., Katsaggelos, A.K., Bagci, U.: Explainable Transformer Prototypes for Medical Diagnoses. arXiv preprint arXiv:2403.06961 (2024)
3. Somani, R. et al.: Explain with Confidence: Fusing Saliency Maps for Faithful and Interpretable Explanations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2025)
4. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)
5. Long, B., Liu, E., Qiu, R., Duan, Y.: Explainable AI - The Latest Advancements and New Trends. arXiv preprint arXiv:2505.07005 (2025)
6. Veeramani, N., Reshma Sherine, S.A., Sakthi Prabha, S., Srinidhi, S., Jayaraman, P.: NextGen Lung Disease Diagnosis with Explainable Artificial Intelligence. Scientific Reports **15**, 33052 (2025)
7. Ullah, N., Guzmán-Aroca, F., Martínez-Álvarez, F., De Falco, I., Sannino, G.: A Novel Explainable AI Framework for Medical Image Classification Integrating Statistical, Visual, and Rule-Based Methods. Medical Image Analysis **105**, 103665 (2025)
8. Aasem, M., Iqbal, M.J.: Toward Explainable AI in Radiology: Ensemble-CAM for Effective Thoracic Disease Localization in Chest X-ray Images Using Weak Supervised Learning. Frontiers in Big Data **7**, 1366415 (2024)
9. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018)
10. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
11. Lundberg, S.M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems, vol. 30, pp. 4768–4777 (2017)
12. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
13. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 3319–3328 (2017)
14. Chowdhury, M.E.H., Rahman, T., Khandakar, A. et al.: Can AI Help in Screening Viral and COVID-19 Pneumonia? IEEE Access **8**, 132665–132676 (2020)
15. Rahman, T., Kiranyaz, S., Al Maadeed, S. et al.: Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection Using Chest X-ray Images. Computers in Biology and Medicine **132**, 104319 (2021)
16. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2097–2106 (2017)
17. Adebayo, J., Gilmer, J., Goodfellow, I., Kim, B.: Sanity Checks for Saliency Maps. In: Advances in Neural Information Processing Systems, vol. 31, pp. 9505–9515 (2018)
18. Chefer, H., Gur, S., Wolf, L.: Transformer Interpretability Beyond Attention Visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 782–791 (2021)
19. Tjoa, E., Guan, C.: A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems **32**(11), 4793–4813 (2020)
20. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R.: Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Transactions on Neural Networks and Learning Systems **28**(11), 2660–2673 (2017)
21. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)
24. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986 (2022)
25. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021)
26. Kasmi, G., Brunetto, A., Fel, T., Parekh, J.: One Wave to Explain Them All: A Unifying Perspective on Post-hoc Explainability. arXiv preprint arXiv:2410.01482 (2024)
27. Mamalakis, A. et al.: Solving the Enigma: Enhancing Faithfulness and Comprehensibility in Explanations of Deep Networks. arXiv preprint arXiv:2405.10008 (2024)
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
29. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 568–578 (2021)
30. Wienholt, L. et al.: MedicalPatchNet: A Patch-Based Self-Explainable AI Architecture for Chest X-ray Classification. arXiv preprint arXiv:2509.07477 (2025)