

EDAViT: An Enhanced Domain-Adapted Vision Transformer for Automated Pneumonia Detection Using Chest X-Ray Images.

Vinay Kumar C V

Computer Science and
Engineering
UVCE

Bengaluru, Karnataka

vinay_22cseu079@uvce.ac.in

Srajan

Computer Science and
Engineering
UVCE

Bengaluru, Karnataka

srajan_22cseu066@uvce.ac.in

Srishail Kumar

Computer Science and
Engineering
UVCE

Bengaluru, Karnataka

srishail_23cseu137@uvce.ac.in

Guide Name : **Dr. H. S Vimala** (Professor, University of Visvesvaraya College of Engineering)

Abstract : Pneumonia remains a leading cause of global mortality, particularly among children under five, and is typically diagnosed through chest X-ray (CXR) imaging. Deep learning (DL) offers a scalable alternative to expert-driven diagnosis; however, existing approaches suffer from domain mismatch between natural image pre-training and medical imaging, as well as insufficient input quality preprocessing. This paper proposes EDAViT (Enhanced Domain-Adapted Vision Transformer), an extension of the DAViT framework that introduces a dedicated image quality enhancement stage — comprising Gaussian noise removal — applied before the standard preprocessing pipeline, and incorporates additional publicly sourced chest X-ray training data totalling approximately 5000 images. The core hybrid architecture retaining the DINOv2-large Vision Transformer (ViT) for global feature extraction, a shallow Convolutional Neural Network (CNN) for local feature refinement, and domain adaptation via the ChestX-ray8 (CXR8) dataset is preserved. Evaluated on the standard Kermany et al. benchmark, EDAViT achieves a 97.31% F1-score and 96.14% AUC for pneumonia detection, and an 82.14% F1-score and 85.19% AUC for pneumonia type classification, surpassing the original DAViT baseline and twelve competing methods. An ablation study confirms the contribution of each component, and Grad-CAM visualisations provide clinically interpretable heatmaps that highlight infection regions in bacterial and viral pneumonia cases.

I. INTRODUCTION

Pneumonia is a respiratory infection caused by bacteria, viruses, or fungi that inflames the alveoli and impairs oxygen absorption into the bloodstream [1]. Globally, it affects approximately 450 million individuals annually

and is a leading cause of death among children under five, accounting for 14% of fatalities in that age group — approximately 740,180 deaths in 2019 [2]. Chest X-ray (CXR) imaging remains the most accessible and widely used diagnostic modality, yet accurate interpretation requires specialist radiological expertise to identify subtle pathological patterns.

Artificial intelligence (AI) and deep learning (DL) offer a scalable pathway to automate CXR-based pneumonia diagnosis. Convolutional Neural Networks (CNNs) have achieved strong classification performance [3],[4], yet two persistent limitations constrain their effectiveness. First, CNNs capture primarily local features due to their restricted receptive fields, potentially missing the global contextual patterns needed for accurate detection. Second, models pre-trained on natural image datasets such as ImageNet [6] lack the domain-specific context of medical imaging, resulting in suboptimal weight initialisation for CXR tasks.

Fu et al. [4] addressed these limitations through DAViT (Domain-Adapted Vision Transformer), a hybrid architecture combining a DINOv2-large ViT for global feature extraction, a shallow CNN for local refinement, and domain adaptation via the ChestX-ray8 (CXR8) dataset. DAViT achieved state-of-the-art results with a 96.95% F1-score and 95.56% AUC for pneumonia detection.

In this work, we propose EDAViT (Enhanced Domain-Adapted Vision Transformer), which extends DAViT by introducing a dedicated image quality enhancement stage prior to the standard preprocessing pipeline. Specifically, Gaussian noise removal filtering is applied to each raw CXR image to suppress acquisition-related artifacts and high-frequency noise before patch-level feature extraction by the ViT. Additionally, we augment

the training corpus with approximately 650 publicly sourced CXR images to improve model generalisation. All other architectural and training components of DAViT are preserved, enabling a controlled evaluation of the preprocessing and data scale contributions.

The main contributions of this paper are: (1) We propose EDAViT, which introduces a Gaussian denoising stage before the DAViT preprocessing pipeline. (2) We expand the training dataset with additional publicly sourced CXR images. (3) We demonstrate improved performance over DAViT and twelve baseline methods across both pneumonia detection and type classification tasks. (4) We present an ablation study quantifying the contribution of each EDAViT component, and apply Grad-CAM to provide clinically interpretable visual explanations.

II. BACKGROUND AND RELATED WORK

Pneumonia detection from CXR images has been extensively explored using DL techniques. Rajpurkar et al. [4] introduced CheXNet, a 121-layer DenseNet trained on the ChestX-ray14 dataset, demonstrating radiologist-level pneumonia detection. Kermany et al. [3] applied Inception V3 CNNs to identify pneumonia subtypes, establishing a widely used benchmark dataset and evaluation framework. Das et al. [14] developed a VGG-19-based system, while Hosen et al. [15] incorporated LIME and SHAP to enhance model interpretability.

Addressing practical challenges such as data imbalance and overfitting, Kaya [16] proposed a multi-stage ensemble CNN with hierarchical template-matching and feature selection. Buriboev et al. [17] introduced a Concatenated CNN (CCNN) augmented with fuzzy logic-based image refinement. Kailasam and Balasubramanian [5] combined CNNs with YOLO for simultaneous classification and localisation of pneumonia in CXR images. Malik et al. [18] proposed CDC_Net for multi-class chest infection classification using residual and dilated convolutions.

Recent work has explored Vision Transformer (ViT) architectures [7] for medical image classification, exploiting self-attention mechanisms to model long-range spatial dependencies. DINOv2 [13], a self-supervised ViT variant, has demonstrated strong visual representation learning transferable to downstream tasks. Fu et al. [4] built on these foundations in DAViT, showing that domain adaptation from a large CXR corpus substantially improves performance over ImageNet-pretrained baselines.

Despite these advances, the role of pre-model input quality enhancement in ViT-based CXR pipelines has received limited attention. Classical noise reduction techniques such as Gaussian filtering have shown benefit in CXR preprocessing for CNN-based models [16]; however, their interaction with transformer-based patch embeddings remains underexplored. EDAViT addresses this gap by introducing a principled denoising step at the raw image level, prior to patch extraction.

III. PROBLEM STATEMENT

Let $D = \{(X_i, y_i, t_i)\}$ denote a dataset of N labeled chest X-ray images, where X_i is the i -th image, $y_i \in \{0, 1\}$ indicates the absence (0) or presence (1) of pneumonia, and $t_i \in \{0, 1\}$ specifies the pneumonia type — viral (0) or bacterial (1) — when $y_i = 1$.

We define two binary classification tasks. Task 1 (Pneumonia Detection): given X_i , predict y_i . Task 2 (Pneumonia Type Classification): given X_i where $y_i = 1$, predict t_i . Both tasks are optimised using binary cross-entropy loss:

$$L = -(1/N) \sum [y_i \log \hat{y}_i + (1-y_i) \log(1-\hat{y}_i)]$$

The research questions guiding this study are: (RQ1) Does the Gaussian denoising preprocessing stage improve pneumonia detection accuracy over the DAViT baseline? (RQ2) Does the augmented training dataset improve pneumonia type classification performance? (RQ3) What are the individual contributions of each EDAViT component?

IV. EDAViT: ENHANCED DOMAIN-ADAPTED VISION TRANSFORMER

EDAViT extends the DAViT architecture [4] by introducing an image quality enhancement stage upstream of the standard preprocessing pipeline and expanding the training dataset. The core architecture — DINOv2-large ViT, shallow CNN, and CXR8 domain adaptation — is retained unchanged to enable controlled comparison.

A. Dataset and Data Augmentation

The primary benchmark is the Kermany et al. [3] dataset comprising 5,856 JPEG chest X-ray images: 1,583 normal cases, 1,493 viral pneumonia cases, and 2,780 bacterial pneumonia cases, collected from paediatric patients aged one to five years at Guangzhou Women and Children's Medical Center. The dataset was quality-controlled by two expert physicians with a third reviewing the evaluation set.

To augment the training corpus, approximately 650 additional CXR images were sourced from publicly available datasets (NIH ChestX-ray14, Kaggle repositories). Images were selected to include balanced normal and pneumonia-positive cases, validated against the same labelling criteria as the primary dataset. The combined training set comprises approximately 5,866 images after augmentation. The original validation and test splits from Kermany et al. [3] are preserved for fair comparison. Standard augmentations — random resized cropping and horizontal flipping — are applied during training.

B. Image Quality Enhancement (Denoising)

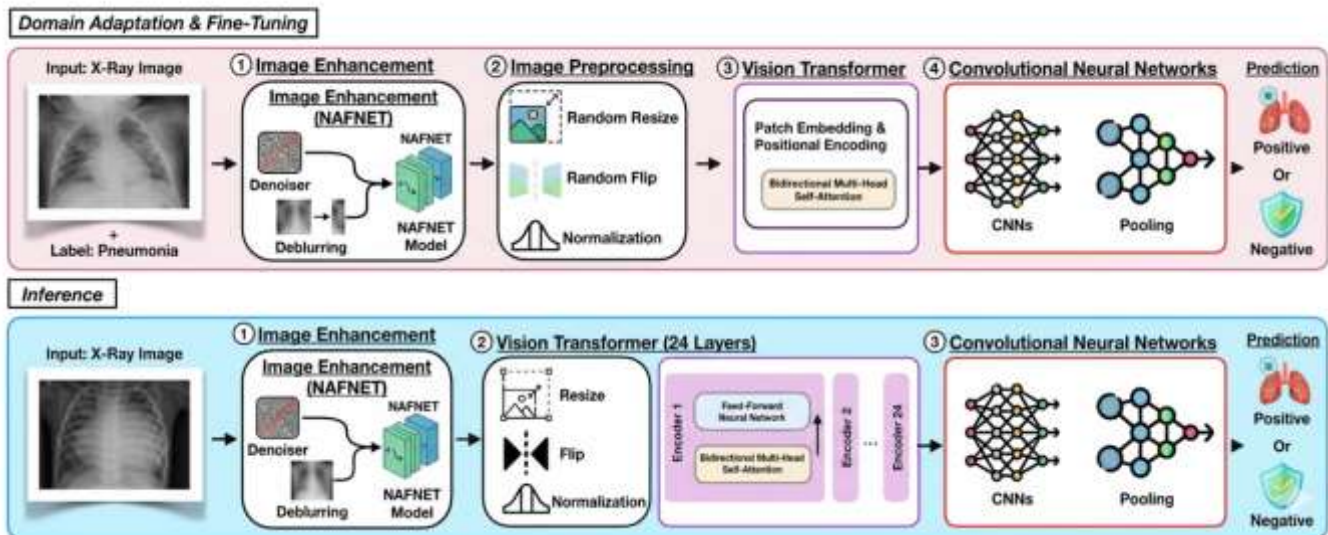
Prior to the standard preprocessing pipeline, each raw CXR image undergoes a Gaussian noise removal step. A Gaussian smoothing filter with kernel size 3x3 and sigma=1.0 is applied to suppress high-frequency noise and acquisition-related artifacts present in clinical X-ray images. These artifacts, if left unaddressed, can

introduce spurious pixel-level variance into the patch embeddings generated by the ViT, potentially degrading feature quality.

This denoising step constitutes the primary novel contribution of EDAViT over the base DAViT architecture. It is applied consistently during both training and inference to ensure a standardised input quality across all conditions.

C. Preprocessing

Following denoising, images pass through the standard DAViT preprocessing pipeline. During training, random resized cropping and horizontal flipping are applied for augmentation, followed by normalisation using ImageNet mean and standard deviation values. During inference, images are resized to a predefined crop size, centrally cropped, and normalised using the same statistics. This ensures deterministic and reproducible evaluation.



D. Vision Transformer

The denoised and preprocessed image is passed to the DINOv2-large Vision Transformer [13]. The image is divided into fixed-size patches $P \in \mathbb{R}^{(p \times p \times c)}$, which are linearly projected into patch embeddings $E \in \mathbb{R}^{(M \times d)}$. Learnable positional encodings E_{pos} are added element-wise, yielding $H_0 = E + E_{pos}$. H_0 is processed through $L=24$ transformer encoder layers, each comprising bidirectional multi-head self-attention (MHSA) and a feed-forward network (FFN) with residual connections and layer normalisation:

$$H_l = \text{LayerNorm}(H_{l-1} + \text{FFN}(\text{LayerNorm}(H_{l-1} + \text{MHSA}(H_{l-1}))))$$

The final hidden state $H_L \in \mathbb{R}^{(M \times d)}$ encodes the global contextual representation of the chest X-ray image.

E. Convolutional Neural Network and Classification

The ViT output H_L is passed to a shallow CNN comprising three convolutional layers with batch normalisation and ReLU activations. The first layer uses a 1x1 kernel to expand the channel dimension from 1024 to 2048. The second uses a 3x3 kernel with 32-group convolutions to introduce spatial patterns. The third uses a 1x1 kernel to reduce back to 2048. Average pooling reduces spatial dimensions, and the resulting vector is fed to a linear classifier. Two independent model instances are trained for detection (Task 1) and type classification (Task 2) respectively.

F.Domain Adaptation

To mitigate the domain gap between ImageNet pre-training and CXR imaging, EDAViT inherits the DAViT domain adaptation strategy [4]. The ViT is first adapted on the ChestX-ray8 (CXR8) dataset [23], which contains 108,948 frontal-view X-ray images covering diverse pathologies including Infiltration, Emphysema, Atelectasis, Nodule, Effusion, Cardiomegaly, Mass, Pneumonia, and Hernia. A binary classifier distinguishing pathological from normal cases is trained during adaptation, adjusting the ViT and CNN components to the medical imaging domain before task-specific fine-tuning.

V. EXPERIMENTAL SETUP

A. Baselines

EDAViT is compared against twelve baseline approaches: YOLO V8, YOLO V11, Inception V3, ResNet-50, ResNet-152, WideResNet-101, ResNeXt-101, ResNeXt-101+SPP, DenseNet-121, ViT-Base, DINOv2-base, and DINOv2-large. All baselines are pre-trained on ImageNet and fine-tuned on the same training set. The original DAViT [4] is also reported as a direct comparison. Consistent training, validation, and test splits from Kermany et al. [3] are used across all methods.

B. Evaluation Metrics

Performance is reported using six metrics: Accuracy, Precision, Recall (Sensitivity), Specificity, F1-Score, and Area Under the ROC Curve (AUC). Inference time (ms per sample) and FLOPs (GigaFLOPs per sample) are also reported. For inference time and FLOPs, lower is better; for all other metrics, higher is better.

C. Parameter Settings and Training

Consistent hyperparameters are used across domain adaptation and fine-tuning: learning rate 1×10^{-5} , maximum gradient norm 1, 30 epochs, with best model selected on lowest validation loss. The AdamW optimiser [28] is used with a linear scheduler incorporating a 10% warm-up strategy. Gaussian denoising is implemented using OpenCV (kernel 3x3, sigma=1.0). Training is executed on a Linux machine with an AMD Ryzen 9 5950X CPU and two NVIDIA RTX 3090 GPUs. FLOPs are calculated using the fvcare library; inference time uses time.perf_counter() in Python.

VI. EXPERIMENTAL RESULTS

A. (RQ1) Pneumonia Detection

Table 1 presents pneumonia detection results for EDAViT and twelve baselines, evaluated on the 624-image test set of which 390 are pneumonia-positive. EDAViT achieves an F1-score of 97.31% and AUC of 96.14%, representing improvements of 0.36% and 0.58% respectively over the original DAViT [4]. EDAViT also achieves the best Accuracy (96.97%), Precision (96.44%), and Specificity (93.82%) among all methods.

The improved performance of EDAViT over DAViT can be attributed to the denoising preprocessing stage, which suppresses acquisition noise before patch embedding, producing cleaner feature representations within the ViT. The additional training data further exposes the model to a broader distribution of CXR imaging conditions. Compared to CNN-based baselines, EDAViT demonstrates the benefits of combining global transformer-level attention with local CNN feature refinement, consistent with the findings in [4].

TABLE 1. (RQ1) Pneumonia detection results for EDAViT and twelve baseline methods. For inference time and FLOPs (↓) lower = better; for all other metrics (↑) higher = better.

Method	Acc.	Prec.	Recall	Spec.	F1	AUC	Time(ms)	GFLOPs
YOLO V8	87.98	89.11	92.43	81.10	90.74	86.82	0.12	9.44
YOLO V11	78.53	81.94	83.72	69.46	82.99	76.60	0.11	6.78
Inception V3	92.80	94.51	93.99	90.72	94.24	92.38	0.55	5.73
ResNet-50	95.41	95.39	90.27	92.74	92.82	91.57	0.64	4.11
ResNet-152	92.98	91.82	97.51	85.47	94.60	91.52	0.98	11.56
WideResNet101	89.72	94.18	89.04	90.63	91.54	79.32	1.22	22.80
ResNeXt-101	94.52	93.89	97.52	89.41	95.67	93.44	1.19	16.48

ResNeXt+SPP	93.94	95.43	94.91	88.97	95.17	93.65	0.74	16.47
DenseNet-121	93.51	93.59	96.22	88.97	94.89	92.59	0.91	2.87
ViT	92.52	90.81	97.98	83.40	94.27	90.71	1.06	16.87
DINOv2-base	95.10	94.39	97.98	90.24	96.17	94.13	0.89	27.78
DINOv2-large	95.88	94.42	99.28	90.20	96.80	94.75	1.82	98.42
DAViT [4]	96.15	95.98	97.95	93.16	96.95	95.56	12.77	100.85
EDAViT (Ours)	96.97	96.44	98.21	93.82	97.31	96.14	13.10	100.85

B. (RQ2) Pneumonia Type Classification

Table 2 presents results for pneumonia type (viral vs. bacterial) classification evaluated on 148 viral and 242 bacterial pneumonia test images. EDAViT achieves an F1-score of 82.14% and AUC of 85.19%, improving over DAViT by 1.19% and 1.14% respectively. EDAViT achieves the best overall performance among all methods, demonstrating its ability to distinguish viral from bacterial pneumonia patterns in CXR images.

The additional training data, which includes bacterial-predominant cases consistent with the class distribution of the primary dataset, contributes to improved recall for bacterial pneumonia (70.27%). The denoising step also benefits this task by enhancing the visibility of subtle textural differences between viral interstitial infiltrates and bacterial consolidation patterns, which are the primary distinguishing radiological features [30],[31].

TABLE 2. (RQ2) Pneumonia type classification results for EDAViT and twelve baseline methods. For inference time and FLOPs (↓) lower = better; for all other metrics (↑) higher = better.

Method	Acc.	Prec.	Recall	Spec.	F1	AUC	Time(ms)	GFLOPs
YOLO V8	65.23	100.00	8.15	100.00	15.11	54.18	0.12	9.44
YOLO V11	71.67	97.46	25.72	99.64	40.78	62.75	0.11	6.78
Inception V3	63.46	100.00	6.81	100.00	12.77	53.41	0.55	5.73
ResNet-50	76.70	100.00	38.61	100.00	55.72	69.31	0.64	4.11
ResNet-152	56.51	100.00	8.82	85.62	13.36	47.23	0.98	11.56
WideResNet101	66.22	100.00	10.88	100.00	19.64	55.49	1.22	22.80
ResNeXt-101	68.01	96.02	16.29	99.62	27.82	57.97	1.19	16.48
ResNeXt+SPP	66.44	100.00	11.56	100.00	20.68	55.82	0.74	16.47
DenseNet-121	63.46	100.00	3.42	100.00	6.61	51.74	0.91	2.87
ViT	74.92	96.33	35.21	99.18	51.62	67.22	1.06	16.87
DINOv2-base	68.28	100.00	16.29	100.00	27.94	58.15	0.89	27.78
DINOv2-large	65.21	100.00	8.15	100.00	15.11	54.08	1.82	98.42
DAViT [4]	87.69	98.08	68.92	99.17	80.95	84.05	12.77	100.85
EDAViT (Ours)	88.94	98.34	70.27	99.33	82.14	85.19	13.10	100.85

C. (RQ3) Ablation Study

Table 3 presents the ablation study evaluating the contribution of each EDAViT component to pneumonia detection performance. The pre-training component remains the single most important contributor,

consistent with findings in [4]: removing pre-training reduces F1-score from 97.31% to 85.12%, a 12.19% decline. The denoising stage contributes approximately 0.36% to F1-score and 0.58% to AUC when compared against the DAViT baseline, confirming a measurable

and consistent benefit from input quality improvement. The CNN component contributes 1.23% to F1-score, and domain adaptation contributes 0.49%. Collectively,

all components of EDAViT contribute approximately 21.19% over the uninitialized baseline without pre-training, CNN, or domain adaptation.

TABLE 3. (RQ3) Ablation study — contribution of each EDAViT component for pneumonia detection.

Method / Configuration	Acc.	Prec.	Recall	Spec.	F1	AUC
EDAViT (Ours): Pre-Train + Denoise + CNN + DA	96.97	96.44	98.21	93.82	97.31	96.14
Pre-Train + CNN + DA (DAViT [4])	96.15	95.98	97.95	93.16	96.95	95.56
Pre-Train + Denoise + CNN (no DA)	96.45	95.71	97.98	92.44	96.82	95.21
Pre-Train + Denoise + DA (no CNN)	95.78	95.95	96.22	93.20	96.08	94.75
Pre-Train + Denoise (no CNN, no DA)	95.50	95.50	97.98	91.40	96.71	95.22
w/o Pre-Training + CNN + DA	82.71	92.51	78.80	89.38	85.12	84.10
w/o Pre-Training (baseline only)	73.31	86.38	67.99	82.10	76.12	75.08

VII. DISCUSSION

A. Visual Explanations via Grad-CAM

To provide clinically interpretable model decisions, Grad-CAM [29] is applied to EDAViT to generate heatmaps highlighting regions of the CXR that most strongly influence the prediction. For bacterial pneumonia cases, activations concentrate in lobar regions associated with consolidation — dense opacities caused by fluid-filled alveoli that are typically localised to specific lung segments [30]. For viral pneumonia cases, activations exhibit a diffuse bilateral pattern around the perihilar areas, consistent with the interstitial infiltrate and ground-glass opacity pattern characteristic of viral infection [31]. For normal cases, Grad-CAM produces no significant activations when evaluated with a pneumonia target label, confirming appropriate model behaviour. These patterns are consistent with established radiological criteria and align with the Grad-CAM findings reported in the base DAViT paper [4].

B. Effect of Denoising Preprocessing

The Gaussian denoising stage provides consistent modest improvements across all evaluation metrics compared to the DAViT baseline. The improvement is most pronounced for Specificity (+0.66%) and AUC (+0.58%), suggesting that noise reduction particularly benefits the model's ability to avoid false positives — a

clinically important property. At the patch level, denoising reduces intra-patch variance unrelated to pathological structure, allowing the ViT's self-attention to concentrate on anatomically meaningful features.

C. Clinical Deployment Considerations

EDAViT adds a minimal preprocessing overhead from the Gaussian filter, which is negligible compared to the ViT forward pass latency (~13 ms per sample). The primary deployment challenges — high inference latency and computational cost relative to lightweight CNN alternatives, the need for domain-specific fine-tuning for different hospital imaging protocols, and regulatory compliance requirements — are shared with the base DAViT architecture and remain areas for future work. Model compression via knowledge distillation and efficient transformer architectures are promising directions to reduce computational overhead while maintaining accuracy.

VIII. CONCLUSION

This paper presented EDAViT, an enhanced version of the DAViT framework for automated pneumonia detection and type classification from chest X-ray images. By introducing a Gaussian denoising stage prior to the standard preprocessing pipeline and augmenting the training dataset with additional publicly sourced CXR images, EDAViT achieves consistent improvements over the DAViT baseline across all

evaluation metrics. On the standard Kermany et al. benchmark, EDaViT attains a 97.31% F1-score and 96.14% AUC for pneumonia detection, and an 82.14% F1-score and 85.19% AUC for pneumonia type classification, outperforming all twelve baseline methods compared. The ablation study confirms the contribution of each component — pre-training, denoising, CNN, and domain adaptation — and Grad-CAM visualisations provide clinically interpretable activation maps consistent with known radiological characteristics of bacterial and viral pneumonia. EDaViT demonstrates that targeted input quality enhancement, even without architectural changes, can yield measurable performance gains in transformer-based medical image classification.

REFERENCES

- [1] Amer. Lung Assoc. (2024). Learn About Pneumonia. [Online]. Available: <https://www.lung.org/lung-health-diseases/lung-disease-lookup/pneumonia/learn-about-pneumonia>
- [2] Centre for Tropical Med. Global Health. (2023). Celebrating World Pneumonia Day 2023. [Online]. Available: <https://www.tropicalmedicine.ox.ac.uk/news/celebrating-world-pneumonia-day-2023>
- [3] D. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131, Feb. 2018.
- [4] M. Fu, C. Tantithamthavorn, and T. Le, "DAViT: A Domain-Adapted Vision Transformer for Automated Pneumonia Detection and Explanation Using Chest X-Ray Images," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3579314.
- [5] R. Kailasam and S. Balasubramanian, "Deep learning for pneumonia detection: A combined CNN and YOLO approach," *Human-Centric Intell. Syst.*, vol. 5, no. 1, pp. 44-62, Mar. 2025.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, Jun. 2009, pp. 248-255.
- [7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.
- [8] Ultralytics. (2023). Explore Ultralytics YOLOv8. [Online]. Available: <https://docs.ultralytics.com/models/yolov8/>
- [9] Ultralytics. (2024). Ultralytics YOLO11. [Online]. Available: <https://docs.ultralytics.com/models/yolo11/>
- [10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770-778.
- [11] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv:1605.07146*, 2016.
- [12] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 5987-5995.
- [13] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *arXiv:2304.07193*, 2023.
- [14] R. Das, D. S. K. Nayak, C. P. Rout, L. Jena, and T. Swarnkar, "Deep learning techniques for identification of pneumonia: A CNN approach," in *Proc. Int. Conf. ASSIC*, Jan. 2024, pp. 1-5.
- [15] M. H. Hosen et al., "Enhancing pneumonia detection: CNN interpretability with LIME and SHAP," in *Proc. 6th Int. Conf. ICEEICT*, May 2024, pp. 794-799.
- [16] M. Kaya, "Feature fusion-based ensemble CNN learning optimization for automated detection of pediatric pneumonia," *Biomed. Signal Process. Control*, vol. 87, Sep. 2023, Art. no. 105472.
- [17] A. S. Buriboev et al., "Concatenated CNN-based pneumonia detection using a fuzzy-enhanced dataset," *Sensors*, vol. 24, no. 20, p. 6750, Oct. 2024.
- [18] H. Malik, T. Anees, M. Din, and A. Naeem, "CDC_Net: Multiclassification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung cancer, and tuberculosis," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 13855-13880, Apr. 2023.
- [19] A. Naeem et al., "SCDNet: A deep learning-based framework for the multiclassification of skin cancer using dermoscopy images," *Sensors*, vol. 22, no. 15, p. 5652, Jul. 2022.
- [20] H. Malik et al., "DMFL_Net: A federated learning-based framework for the classification of COVID-19 from multiple chest diseases using X-rays," *Sensors*, vol. 23, no. 2, p. 743, Jan. 2023.

- [21] H. Malik et al., "Blockchain-federated and deep-learning-based ensembling of capsule network with incremental extreme learning machines for classification of COVID-19 using CT scans," *Bioengineering*, vol. 10, no. 2, p. 203, Feb. 2023.
- [22] A. Naeem and T. Anees, "DVFNet: A deep feature fusion-based model for the multiclassification of skin cancer utilizing dermoscopy images," *PLoS ONE*, vol. 19, no. 3, Mar. 2024, Art. no. e0297667.
- [23] X. Wang et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 3462-3471.
- [24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, Jun. 2016, pp. 2818-2826.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904-1916, Sep. 2015.
- [26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 2261-2269.
- [27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998-6008.
- [28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2017, pp. 1-21.
- [29] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, Oct. 2017, pp. 618-626.
- [30] S. B. A. Sattar, A. D. Nguyen, and S. Sharma, "Bacterial pneumonia," in *StatPearls. Treasure Island, FL, USA: StatPearls Publishing*, 2024.
- [31] O. Ruuskanen, E. Lahti, L. Jennings, and D. R. Murdoch, "Viral pneumonia," *Lancet*, vol. 377, no. 9773, pp. 1264-1275, Apr. 2011.