IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2025, accepted 10 June 2025, date of publication 12 June 2025, date of current version 19 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3579314

RESEARCH ARTICLE

DAViT: A Domain-Adapted Vision Transformer for Automated Pneumonia Detection and Explanation Using Chest X-Ray Images

MICHAEL FU[1], CHAKKRIT TANTITHAMTHAVORN[2], (Senior Member, IEEE), AND TRUNG LE[2], (Member, IEEE)

[1]School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC 3000, Australia
[2]Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia

Corresponding author: Michael Fu (michael.fu@unimelb.edu.au)

This work was supported in part by the Faculty of Information Technology (FIT) Early Career Researcher Seed Grant 2024 from Monash University, Australia.

ABSTRACT Pneumonia, a leading cause of mortality worldwide, especially among children under five, is typically diagnosed via chest X-rays. However, detecting it is challenging as expert radiologists must discern subtle patterns. Artificial intelligence (AI) offers a scalable alternative by automating diagnosis through deep learning (DL) models. Despite progress, current methods face two key limitations: 1) reliance on CNNs that capture local but may overlook global features, and 2) the use of pre-trained models from natural image datasets like ImageNet, which lack the contextual relevance of medical imaging, leading to suboptimal performance. To address these challenges, we propose DAViT (Domain-Adapted Vision Transformer), a hybrid architecture that combines Vision Transformers (ViTs) and shallow CNNs with domain adaptation. The ViT leverages self-attention to capture global features, while the CNN extracts local ones. To mitigate domain differences, we adapt the model using a diverse chest X-ray dataset. We evaluate DAViT on a real-world dataset of 5,856 chest X-rays. The results demonstrate that DAViT achieves state-of-the-art performance with a 97% F1-score and 96% AUC for pneumonia detection, outperforming twelve baseline methods. For pneumonia type classification, DAViT achieves an 81% F1-score and 84% AUC, outperforming baselines by 25% to 74%. An ablation study highlights the critical contributions of domain adaptation, ViT, and CNN components, collectively enhancing performance by 21%. Finally, we apply Grad-CAM on top of DAViT to generate interpretable heatmaps that highlight relevant areas for bacterial and viral pneumonia cases, providing insights to assist medical practitioners in decision-making. These findings indicate the potential of DAViT to assist clinicians in pneumonia diagnosis through improved model accuracy and interpretability. The training code and pre-trained models are available at https://github.com/awsm-research/DAViT

INDEX TERMS Domain adaptation in vision transformers, deep learning for chest X-ray images, explainable AI for X-ray images, pneumonia detection, pneumonia type classification.

I. INTRODUCTION

Pneumonia is a lung infection caused by bacteria, viruses, or fungi, resulting in inflammation of the air sacs (alveoli), which fill with fluid or pus and hinder oxygen absorption into the bloodstream [1]. Globally, it affects approximately 450 million people annually and remains a leading cause of mortality among children under five, accounting for 14% of deaths in this age group, or 740,180 fatalities in 2019 [2]. Chest X-rays are currently the most effective diagnostic tool for pneumonia. However, accurate detection of pneumonia in chest X-rays presents significant challenges, requiring the

The associate editor coordinating the review of this manuscript and approving it for publication was Chih-Yu Hsu.

expertise of trained radiologists to identify subtle patterns indicative of infection.

Artificial intelligence (AI) offers a scalable solution for automated pneumonia diagnosis by leveraging deep learning (DL) models to classify chest X-ray images, enabling rapid analysis of large datasets. Kermany et al. [3] demonstrated the effectiveness of convolutional neural networks (CNNs) such as Inception V3 for detecting pneumonia and explaining its types. Similarly, Rajpurkar [4] employed the DenseNet CNN architecture for automated pneumonia detection without reliance on expert radiologists.

Despite these advancements in DL-based pneumonia detection, we still found two major limitations. First, CNN architectures used in prior works [3], [4], [5] are limited to capturing local features within an image due to their small receptive fields, potentially missing global patterns crucial for accurate diagnosis. Second, existing models are often pre-trained on ImageNet [6], a dataset of natural images that lack medical imaging context, leading to suboptimal weight initialization for tasks involving chest X-rays due to significant domain differences.

In this paper, we propose DAViT (Domain-Adapted Vision Transformer), a hybrid architecture combining a vision transformer (ViT) and a shallow CNN with domain adaptation, to address the identified limitations. The ViT component, with its self-attention mechanisms to capture global features, has outperformed CNN-based architectures on several image benchmark datasets [7]. Meanwhile, the shallow CNN complements the ViT by extracting fine-grained local features. To mitigate the domain difference between natural images and chest X-rays, we employ domain adaptation using a large, comprehensive chest X-ray dataset containing diverse pathological conditions. Finally, we conduct an experiment to compare our DAViT approach with two state-of-the-art (SOTA) CNN-based methods for pneumonia detection and classification, which use Inception V3 [3] and DenseNet [4]. Additionally, we include other SOTA CNN architectures such as YOLO [8], [9], ResNet [10], WideResNet [11], and ResNeXt [12]. We also incorporate vision transformer (ViT)-based architectures, including the original ViT [7] and DINOv2 [13].

A. NOVELTY & CONTRIBUTIONS

To the best of our knowledge, the main contributions of this paper are as follows: (1) We propose DAViT (Domain-Adapted Vision Transformer), a novel hybrid architecture that combines Vision Transformers and shallow CNNs, enhanced with domain adaptation techniques to effectively bridge the gap between natural and medical imaging domains. (2) We conduct a comprehensive evaluation on a real-world dataset of 5,856 chest X-rays. (3) We perform ablation studies and apply Grad-CAM to interpret model predictions, offering insights into the contributions of each component and improving model transparency for clinical support. (4) We publicly release the training code and pre-trained models to promote open science, reproducibility, and future research in AI-driven medical diagnosis.

B. PAPER ORGANIZATION

Section II presents the background and related works. Section III presents the problem statement. Section IV presents our DAViT approach. Section V presents the motivation of our three research questions, our studied datasets, and our experimental setup, while Section VI presents the experimental results. Section VII presents the discussion regarding the interpretability of our DAViT. Section VIII discloses the limitations of this study. Section IX draws the conclusions.

II. BACKGROUND & RELATED WORKS

Pneumonia remains a major global health concern, especially for young children and the elderly. Chest X-ray (CXR) imaging is widely used for diagnosis due to its availability and cost-effectiveness, but interpreting these images can be challenging, even for experienced radiologists. In response, machine learning (ML) and deep learning (DL) have been increasingly applied to automate pneumonia detection. Convolutional neural networks (CNNs) have shown strong performance in extracting visual features from CXR images, enabling more accurate and scalable diagnosis. This has led to a growing body of research focused on improving the accuracy, interpretability, and generalizability of DL-based diagnostic systems.

Pneumonia detection using CXR images has been extensively explored with the advancement of ML and DL techniques. Early work by Rajpurkar [4] introduced CheXNet, a 121-layer CNN trained on the large-scale ChestX-ray14 dataset, demonstrating performance surpassing average radiologists in detecting pneumonia and other thoracic diseases. Kermany et al. [3] further investigated CNNs such as Inception V3 to identify pneumonia and differentiate its subtypes, marking an important step in explainable DL applications in medical imaging. Das et al. [14] utilized a VGG-19-based CNN to develop an autonomous pneumonia detection system, while Hosen et al. [15] incorporated explainability methods like LIME and SHAP to enhance model interpretability for clinical acceptance. Several recent efforts have addressed practical challenges in medical image analysis, including data imbalance, noise, and overfitting. Kaya [16] proposed a multi-stage approach combining hierarchical template-matching for noise reduction, transfer learning with multiple CNNs, and feature selection using Chi-Square and mRMR methods. Buriboev et al. [17] introduced a Concatenated CNN (CCNN) architecture enhanced with a fuzzy logic-based image refinement technique to boost feature extraction quality. More recently, Kailasam and Balasubramanian [5] integrated CNNs with the YOLO algorithm to enable both the classification and localization of pneumonia cases in CXR

images, reflecting growing interest in combining detection with interpretability and localization.

Recent studies also leverage deep learning and federated learning techniques to address challenges in medical image classification, focusing on chest and skin disease detection, privacy preservation, and handling data variability and imbalance. Malik et al. [18] proposes CDC Net, a deep learning-based multi-class classification model designed to identify five major chest infections-COVID-19, lung cancer, pneumothorax, tuberculosis, and pneumonia-from chest X-ray images using a unified approach with residual and dilated convolutions. Naeem et al. [19] presents SCDNet, a framework that integrates Vgg16 with convolutional neural networks for multiclass classification of skin cancer types such as Melanoma, Melanocytic Nevi, Basal Cell Carcinoma, and Benign Keratosis using the ISIC 2019 dataset. Malik et al. [20] introduces DMFL_Net, a federated learning-based framework that utilizes DenseNet-169 to classify COVID-19 and four other chest diseases while preserving data privacy across healthcare institutions. Malik et al. [21] proposes a blockchain-based federated learning framework that combines CapsNet with incremental extreme learning machines to detect COVID-19 from chest CT scans using data from multiple hospitals, addressing privacy and variability challenges. Naeem and Anees [22] introduces DVFNet, a deep learning-based approach combining VGG19 and HOG features for multiclass skin cancer detection from dermoscopy images, incorporating anisotropic diffusion for preprocessing and SMOTE Tomek for class imbalance correction using the ISIC 2019 dataset.

Unlike previous studies that primarily rely on CNN architectures and pretraining on natural images, our work addresses two key challenges in pneumonia detection: limited global feature learning and domain mismatch. We propose a novel hybrid framework, DAViT, that combines Vision Transformers and shallow CNNs to capture both global and local features effectively and leverages domain adaptation techniques to bridge the gap between natural image pretraining and the medical imaging domain.

III. PROBLEM STATEMENT

Let us consider a dataset of N chest X-ray images. The dataset contains images labeled as indicating either pneumonia or not (negative), with additional labels specifying the type of pneumonia as viral or bacterial. We denote an image as $X_i \in \mathcal{X}$, where $i \in \{1, 2, \ldots, N\}$, and let a sample of data be $\{(X_i, y_i, t_i) : y_i \in \mathcal{Y}, t_i \in \mathcal{T}\}$. Here, $\mathcal{X}$ denotes the set of chest X-ray images. $\mathcal{Y} = \{0, 1\}$ represents the presence or absence of pneumonia, where 1 indicates that pneumonia is present and 0 indicates that it is absent. $\mathcal{T} = \{0, 1\}$ specifies the type of pneumonia when present, with 1 representing bacterial and 0 representing viral.

We formulate pneumonia detection as a binary classification task. Given an input X_i, we use a vision transformer (ViT) and a convolutional neural network (CNN) to extract feature embeddings, denoted as $\mathbf{F}_{\text{det}} \in \mathbb{R}^d$, where d is the hidden size. These embeddings are then passed through a linear layer, $\text{Linear}_{\text{det}}$, to predict $y_i \in \mathcal{Y}$. Similarly, we formulate pneumonia type explanation as a second binary classification task. For images where $y_i = 1$, the feature embeddings, denoted as $\mathbf{F}_{\text{exp}} \in \mathbb{R}^d$, are extracted using a separate model with the same architecture. These embeddings are then processed through a linear layer, $\text{Linear}_{\text{exp}}$, to predict $t_i \in \mathcal{T}$, identifying whether the pneumonia is caused by a virus or bacteria.

Our method uses two separate stacks of transformer layers, denoted as $\mathcal{F}_{\text{det}}$ and $\mathcal{F}_{\text{exp}}$, to enhance the feature embeddings for the detection and explanation tasks, respectively. Specifically, the feature embeddings $\mathbf{F}_{\text{det}} \in \mathbb{R}^d$ are processed through $\mathcal{F}_{\text{det}}$, while $\mathbf{F}_{\text{exp}} \in \mathbb{R}^d$ are processed through $\mathcal{F}_{\text{exp}}$. During supervised training, we leverage the labeled dataset to optimize the parameters of $\mathcal{F}_{\text{det}}$, $\mathcal{F}_{\text{exp}}$, and the respective linear classification layers $\text{Linear}_{\text{det}}$ and $\text{Linear}_{\text{exp}}$. For optimization, we minimize the binary cross-entropy loss $\mathcal{L}_{\text{det}}$ for the detection task and $\mathcal{L}_{\text{exp}}$ for the explanation task, defined as:

$$\mathcal{L}_{\text{det}} = -\frac{1}{N_{\text{det}}} \sum_{i=1}^{N_{\text{det}}} \big(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\big),$$

$$\mathcal{L}_{\text{exp}} = -\frac{1}{N_{\text{exp}}} \sum_{i=1}^{N_{\text{exp}}} \big(t_i \log \hat{t}_i + (1 - t_i) \log(1 - \hat{t}_i)\big),$$

where N_{det} and N_{exp} are the numbers of samples for the detection and explanation tasks, respectively, y_i and t_i are the ground truth labels, and $\hat{y}_i$ and $\hat{t}_i$ are the predicted probabilities.

IV. DAViT: A DOMAIN-ADAPTED VISION TRANSFORMER

In this section, we present the design rationale and the architecture of our DAViT approach.

Design Rationale: To address the limitation of CNNs, which primarily capture local features and struggle to encode global contextual information, we design our architecture to incorporate a vision transformer (ViT) component [7] which can effectively capture global features of an image through its self-attention mechanism. We also introduce a shallow CNN to complement our ViT component and focus on fine-grained local feature extraction. To mitigate the domain gap between general image datasets like ImageNet [6] and chest X-ray images, we apply domain adaptation by pre-training our architecture on an extensive chest X-ray dataset. This phase adjusts the pre-trained model parameters to the medical imaging domain, enhancing their relevance for the target task. Following this, the model is then fine-tuned for the final downstream tasks of pneumonia detection and explanation.

Fig 1 presents an overview architecture of our DAViT approach. In what follows, we introduce its technical details.

A. IMAGE PREPROCESSING

In Step ①, we preprocess the chest X-ray images to make them suitable for deep learning models. Specifically, given

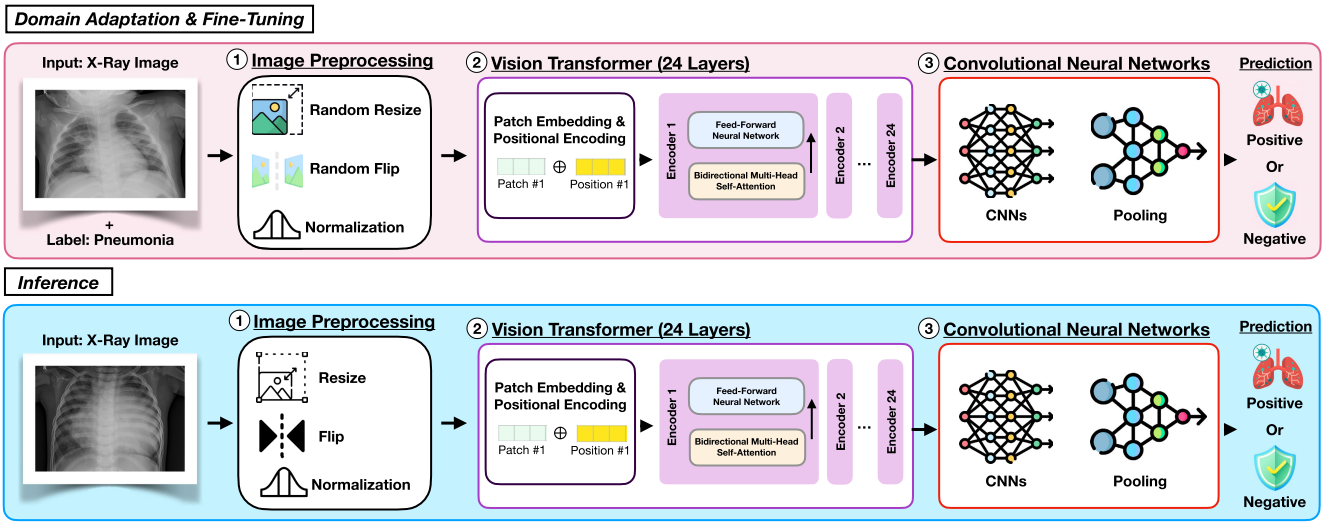

FIGURE 1. An overview architecture of our DAViT approach. The framework includes three main components: (1) Image preprocessing, applying transformations like resizing, flipping, and normalization; (2) Vision transformer (ViT), utilizing patch projection and positional encoding for feature extraction via transformer encoders; and (3) Convolutional neural networks (CNNs) for additional feature refinement and classification.

an input X_i, the image undergoes resizing and cropping to a predefined size, followed by normalization using the mean and standard deviation values derived from the ImageNet dataset [6]. After normalization, the processed image is represented as patches $\mathbf{P} \in \mathbb{R}^{p \times p \times c}$, where p is the patch size, and c is the number of channels (e.g., 3 for RGB images). These patches are subsequently used as input for the vision transformer model.

During training, we apply common transformations used to augment the data and introduce variability. Specifically, we use random resized cropping to extract regions of interest, random horizontal flipping to simulate variations in orientation, and normalization using mean and standard deviation values derived from the ImageNet dataset [6], as provided by the feature processor. These augmentations could help the model generalize better to unseen images by exposing it to a wider range of input conditions.

During inference, we employ a consistent preprocessing pipeline to ensure uniformity and eliminate randomness. The images are resized to a predefined crop size and then centrally cropped to focus on the most relevant regions. Finally, they are normalized using the same ImageNet mean and standard deviation values. This preprocessing pipeline ensures that the input is standardized for the evaluation of the trained model.

B. VISION TRANSFORMER

In Step ②, we use a Vision Transformer (ViT) architecture [7] to extract feature embeddings from the preprocessed chest X-ray images. The patch representation $\mathbf{P} \in \mathbb{R}^{p \times p \times c}$ from Step ① is first flattened and linearly projected into a sequence of embeddings $\mathbf{E} \in \mathbb{R}^{M \times d}$, where M is the number of patches and d is the hidden dimension. To retain positional information, a learnable positional encoding $\mathbf{E}_{pos} \in \mathbb{R}^{M \times d}$ is added element-wise to the patch embeddings, resulting in the input sequence $\mathbf{H}_0 = \mathbf{E} + \mathbf{E}_{pos}$.

The sequence $\mathbf{H}_0$ is then passed through a stack of L transformer encoder layers, each comprising multi-head self-attention and feed-forward networks with residual connections and layer normalization. This process is denoted as:

$$\mathbf{H}_l = \text{LayerNorm}\,(\mathbf{H}_{l-1} + \text{FFN}\,(\text{LayerNorm}\,(\mathbf{H}_{l-1} + \text{MHSA}(\mathbf{H}_{l-1}))))$$

where $\mathbf{H}_{l-1}$ is the input to the l-th layer, $\text{MHSA}(\mathbf{H}_{l-1})$ represents the multi-head self-attention operation on $\mathbf{H}_{l-1}$, $\text{FFN}(\cdot)$ represents the feed-forward network applied to the result of the self-attention, $\text{LayerNorm}(\cdot)$ denotes layer normalization applied to the inputs and outputs of each sub-layer, and $\mathbf{H}_l$ is the output of the l-th layer. This process is repeated through all L layers, with the final output being $\mathbf{H}_L \in \mathbb{R}^{M \times d}$, which encodes the features of the chest X-ray image.

C. CONVOLUTIONAL NEURAL NETWORK

In Step ③, we use a shallow Convolutional Neural Network (CNN) to process the output of the ViT and perform the pneumonia detection and explanation tasks. The last hidden state $\mathbf{H}_L \in \mathbb{R}^{M \times d}$ from the ViT is first passed through the CNN architecture, which consists of three convolutional layers with batch normalization, followed by a ReLU activation. The first convolutional layer uses a 1×1 kernel to increase the channel dimension from 1024 to 2048, while the second convolutional layer uses a 3×3 kernel with grouped convolutions (32 groups) to introduce spatial patterns. The final convolutional layer uses a 1×1 kernel to reduce the channel dimension back to 2048.

After passing through the CNN, the output feature map is pooled using average pooling to reduce the spatial

dimensions. This pooled output is then flattened into a vector to be fed into a fully connected layer. The fully connected layer processes this flattened feature vector, and the final logits are obtained through a linear classifier. This classifier outputs the prediction for pneumonia detection (i.e., indicating the presence of pneumonia) or pneumonia-type explanation (i.e., identifying the type of pneumonia: viral or bacterial).

D. DOMAIN ADAPTATION

We perform domain adaptation to adjust our model to the specific characteristics of chest X-ray images. The ViT in our architecture was pre-trained on the ImageNet dataset [6], which, while extensive, does not include medical images such as chest X-rays. To bridge this gap, we first train our model on the ChestX-ray8 (CXR8) dataset [23], which contains a wide variety of frontal-view chest X-ray images with multiple pathologies, including conditions such as ''Infiltration,'' ''Emphysema,'' ''Atelectasis,'' ''Nodule,'' ''Effusion,'' ''Cardiomegaly,'' ''Mass,'' ''Pneumonia,'' and ''Hernia.''

During domain adaptation, we introduce a classifier designed to identify whether an image contains any pathologies or is benign. This step allows us to adapt the ViT and CNN components of our model to the medical imaging domain, ensuring that the learned representations are more suitable for chest X-ray images. After the domain adaptation phase, we load the pre-trained weights for all components of the model, including the ViT, CNN, and the classifier, and fine-tune the entire architecture for the final downstream pneumonia detection and explanation tasks.

V. EXPERIMENTAL SETUP

In this section, we present the motivation behind our three research questions, provide details about the dataset used in our study along with the data-splitting strategy applied in our experiments, and outline the experimental setup, including baseline approaches, evaluation metrics, parameter settings, and model training procedures.

A. RESEARCH QUESTIONS

To evaluate our DAViT approach, we formulate the following three research questions.

(RQ1) How accurate is our DAViT for pneumonia detection using chest X-ray images? Recently, Kermany et al. [3] and Rajpurkar [4] proposed state-of-the-art approaches for pneumonia detection using convolutional neural networks (CNNs). However, as mentioned in Section I, these methods face two key limitations that affect detection effectiveness. We propose DAViT, designed to overcome these limitations. Thus, we evaluate whether DAViT outperforms these state-of-the-art approaches.

(RQ2) How accurate is our DAViT for explaining pneumonia types using chest X-ray images? Identifying the type of pneumonia, specifically whether it is viral or bacterial, is critical for guiding appropriate treatment strategies, as bacterial pneumonia typically requires antibiotic therapy, whereas viral pneumonia does not. Accurate classification of pneumonia types can therefore improve patient outcomes and reduce unnecessary antibiotic use. Thus, we evaluate the effectiveness of DAViT in explaining pneumonia types compared to existing state-of-the-art approaches.

(RQ3) What are the contributions of the components of our DAViT? Our DAViT approach involves three key components: the Vision Transformer (ViT), the shallow CNN, and the domain adaptation. However, little is known about the individual contributions of these components and which component contributes the most to the accuracy of DAViT. Thus, we formulate this RQ to conduct an ablation study on the variants of DAViT.

B. STUDIED DATASET

To evaluate deep learning (DL)-based models for pneumonia detection and classification, we use the open-source medical dataset compiled by Kermany et al. [3]. This dataset contains 5,856 chest X-ray images in JPEG format, categorized into three groups: 1,583 normal cases, 1,493 pneumonia cases caused by viral infections, and 2,780 pneumonia cases caused by bacterial infections. The chest X-ray images were collected by Kermany et al. [3] from pediatric patients aged one to five years at Guangzhou Women and Children's Medical Center. As described by the original authors, the dataset underwent a rigorous quality control process, where low-quality or unreadable scans were removed, and the diagnoses were assigned by two expert physicians, with a third expert reviewing the evaluation set to ensure labeling accuracy. The datasets used in this study were collected from previously conducted studies in which informed consent had already been obtained. These datasets are publicly available as open-source resources.

The dataset selection was based on several criteria: it is fully open-source with no privacy restrictions, includes high-quality images with verified expert annotations, provides sufficient sample size for deep learning, and is publicly available to support reproducibility. To improve generalization and reduce overfitting, we applied basic data augmentation techniques, including random resizing and horizontal flipping, rather than oversampling. These augmentations help expose the model to a wider variety of image variations without altering the underlying data distribution.

C. DATA SPLITTING

We use the same training, validation, and testing sets provided by Kermany et al. [3]. Specifically, the training set comprises 5,216 X-ray images, of which 3,875 are pneumonia cases. The validation set includes 16 X-ray images with 8 pneumonia cases. The testing set consists of 624 X-ray images, of which 390 are pneumonia cases.

D. BASELINE APPROACHES

To answer RQ1 and RQ2, we compare our DAViT approach with two state-of-the-art (SOTA) CNN-based methods for

pneumonia detection and classification, which utilize Inception V3 [3] and DenseNet [4]. Additionally, we incorporate other SOTA CNN architectures, including YOLO [8], [9], ResNet [10], WideResNet [11], and ResNeXt [12]. Furthermore, we include vision transformer (ViT)-based architectures, such as ViT [7] and DINOv2 [13]. The details of each baseline model are introduced as follows:

- **YOLO** [8], [9]: The YOLO (You Only Look Once) series features advanced backbone and neck architectures for improved feature extraction, an anchor-free Ultralytics head for enhanced accuracy, and an optimized accuracy-speed tradeoff, making it well-suited for real-time detection [8]. We select two of the latest YOLO versions, YOLO V8 and YOLO V11, pre-trained on ImageNet and adaptable to our pneumonia detection task. Specifically, we use two checkpoints, YOLOv8x-cls and YOLOv11x-cls, which have demonstrated the best image-classification performance on ImageNet compared with other YOLO checkpoints [9].
- **Inception V3** [24]: Kermany et al. [3] leveraged this architecture to develop an SOTA method for detecting pneumonia and determining its cause (viral or bacterial). Inception V3 is a deep convolutional neural network (CNN) designed to improve classification accuracy by employing factorized convolutions, auxiliary classifiers, and batch normalization [24]. In our experiments, we consider the ''Inception V3'' model pre-trained on the ImageNet dataset [6].
- **ResNet** [10]: The Residual Network (ResNet) is a deep CNN architecture designed to address the vanishing gradient problem in very deep networks by introducing residual learning through skip (shortcut) connections. These connections allow the network to learn residual mappings instead of direct mappings, facilitating the training of extremely deep models. In our experiments, we consider the commonly used ''ResNet-50'' and the larger ''ResNet-152'', both pre-trained on the ImageNet dataset [6].
- **WideResNet** [11]: Wide Residual Networks (WideResNet) are a variation of ResNet that improves performance by increasing the width (number of feature maps) of residual blocks rather than their depth like original ResNet. In our experiments, we consider ''WideResNet-101'', a commonly used model pre-trained on the ImageNet dataset [6].
- **ResNeXt** [12]: ResNeXt is a deep CNN that extends ResNet by introducing a cardinality dimension, which represents the number of parallel paths (or groups) within each residual block. This design increases the capacity and diversity of the model without significantly increasing computational complexity. In our experiments, we consider ''ResNeXt-101'', a widely used model pre-trained on the ImageNet dataset [6]. Additionally, we include a variant with Spatial Pyramid Pooling (SPP) [25] to enhance global feature capture by aggregating multi-scale spatial information before classification.
- **DenseNet** [26]: Rajpurkar [4] leveraged ''DenseNet-121'' to develop a SOTA pneumonia detection model with promising performance. DenseNet (Densely Connected Convolutional Network) is a model architecture that connects each layer to every other layer in a feed-forward manner, promoting feature reuse and alleviating the vanishing gradient problem when training a deep model. In our experiments, we consider ''DenseNet-121'', commonly pre-trained on the ImageNet dataset [6].
- **ViT** [7]: The Vision Transformer (ViT) is an architecture that applies the transformer model [27], originally designed for natural language processing, to image classification tasks. It divides an image into fixed-size patches, embeds them as input tokens, and processes them using self-attention mechanisms, enabling the model to capture global contextual information. In our experiments, we consider ''ViT-Base'', a commonly used model pre-trained on the ImageNet dataset [6].
- **DINOv2** [13]: DINOv2 (Self-Distillation with No Labels v2) is a self-supervised vision transformer model designed to learn high-level image representations without requiring labeled data. It leverages self-distillation, where a student model learns from a teacher model, enabling effective feature learning across diverse datasets. In our experiments, we consider ''DINOv2-base'' and ''DINOv2-large'', pre-trained on large-scale datasets such as ImageNet.

E. EVALUATION METRICS

Similar to previous studies on binary pneumonia detection and binary pneumonia type classification (i.e., viral vs. bacterial) [3], [23], we report six evaluation metrics as follows:

- **Accuracy:** Measures the overall correctness of the model by evaluating the proportion of correctly classified instances (both positive and negative).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precision:** Evaluates the proportion of true positive cases among all predicted positive cases, reflecting the model's ability to avoid false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall (Sensitivity):** Measures the proportion of actual positive cases correctly identified by the model, assessing its ability to detect pneumonia cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Specificity:** Calculates the proportion of actual negative cases correctly identified, indicating the model's ability

to avoid false alarms.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- **F1-Score:** Provides a harmonic mean of precision and recall, offering a balanced metric when the dataset has imbalanced classes.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Area Under Curve (AUC):** Represents the area under the Receiver Operating Characteristic (ROC) curve, quantifying the model's ability to distinguish between positive and negative classes across various thresholds.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

where the True Positive Rate (TPR) is defined as $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ and the False Positive Rate (FPR) is defined as $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$.

To evaluate the efficiency of each baseline, we measure both inference time and floating-point operations (FLOPs). Inference time reflects the real-time performance of the model, while FLOPs provide an estimate of the computing power required for a single forward pass, helping assess the trade-off between accuracy and efficiency.

- **Inference Time:** We measure inference time to evaluate how quickly each model processes an input sample. This is computed using the `time.perf_counter()` function in Python and is reported in milliseconds (ms) per sample.
- **FLOPs:** Floating-point operations (FLOPs) indicate the number of arithmetic operations required for a single forward pass. A higher FLOP count generally corresponds to greater computing power demand. We calculate FLOPs using the `fvcore` Python library and report the results in GigaFLOPs (GFLOPs) per sample.

F. PARAMETER SETTINGS AND MODEL TRAINING

We use the CXR8 (ChestX-ray8) dataset, curated by Wang et al. [23], for domain adaptation. This dataset comprises 108,948 frontal-view X-ray images, with 24,636 images labeled as containing one or more pathologies, including conditions such as "Infiltration," "Emphysema," "Atelectasis," "Nodule," "Effusion," "Cardiomegaly," "Mass," "Pneumonia," and "Hernia." The remaining 84,312 images represent normal cases. This large-scale dataset provided a comprehensive foundation for adapting our models to X-ray images.

We employ consistent parameter settings for domain adaptation and fine-tuning to address the downstream pneumonia detection and explanation tasks. During training, we use a learning rate of 1×10^{-5}, a maximum gradient norm of 1, and 30 epochs, with the best model selected based on the lowest validation loss. We use the AdamW optimizer [28], with a linear scheduler incorporating a 10% warm-up step strategy.

Model training is executed on a Linux machine equipped with an AMD Ryzen9 5950X CPU and two NVIDIA RTX 3090 GPUs. The full hyperparameter settings and training recipe are open-sourced at https://github.com/awsm-research/DAViT

VI. EXPERIMENTAL RESULTS

In this section, we present the results of our three research questions. Figure 2 presents an overview of our experimental design and results.

A. (RQ1) HOW ACCURATE IS OUR DAViT FOR PNEUMONIA DETECTION USING CHEST X-RAY IMAGES?

Approach. To address this RQ, we evaluate the accuracy of DAViT for pneumonia detection using chest X-ray images. We compare DAViT with twelve baseline approaches on the X-ray image dataset for pneumonia detection collected by Kermany et al. [3]. Specifically, we train each model using the training set, select the best model based on the validation loss for each approach, and evaluate performance using the same testing set, which consists of 624 X-ray images, 390 of which indicate pneumonia.

Results. Table 1 presents the experimental results for pneumonia detection by DAViT and the twelve baseline approaches, evaluated using six metrics: accuracy, precision, recall, specificity, F1-score, AUC, per-sample inference time, and FLOPs.

Our DAViT demonstrates the best overall performance, achieving the highest F1-score of 97% and an AUC of 96%, effectively balancing accurate pneumonia detection with less false positives. Furthermore, our DAViT also achieves the best accuracy of 96%, precision of 96%, and specificity of 93%. When comparing DAViT with state-of-the-art pneumonia detection methods, our approach demonstrates substantial improvements. In particular, it outperforms the Inception V3 model used by Kermany et al. [3] and the DenseNet-121 model employed by Rajpurkar [4]. In addition, applying the Spatial Pyramid Pooling (SPP) layer on top of ResNeXt-101 achieves performance comparable to the original ResNeXt-101. This suggests that global feature extraction has a limited effect on ResNet-based CNNs in the context of pneumonia detection. In terms of F1-score, DAViT achieves an improvement of 3% and 2%, respectively, over these models. Similarly, it delivers a 4% and 3% enhancement in AUC. These results confirm that our DAViT approach can detect pneumonia using chest X-ray images more effectively than existing deep learning-based approaches.

The improved performance of DAViT may be attributed to its hybrid architecture, which combines convolutional operations with vision transformer-based attention mechanisms. This design enables the model to capture both local features and global contextual information from chest X-ray images, which can aid in distinguishing pneumonia from normal cases. Additionally, the sequential pipeline—where global representations from the transformer are

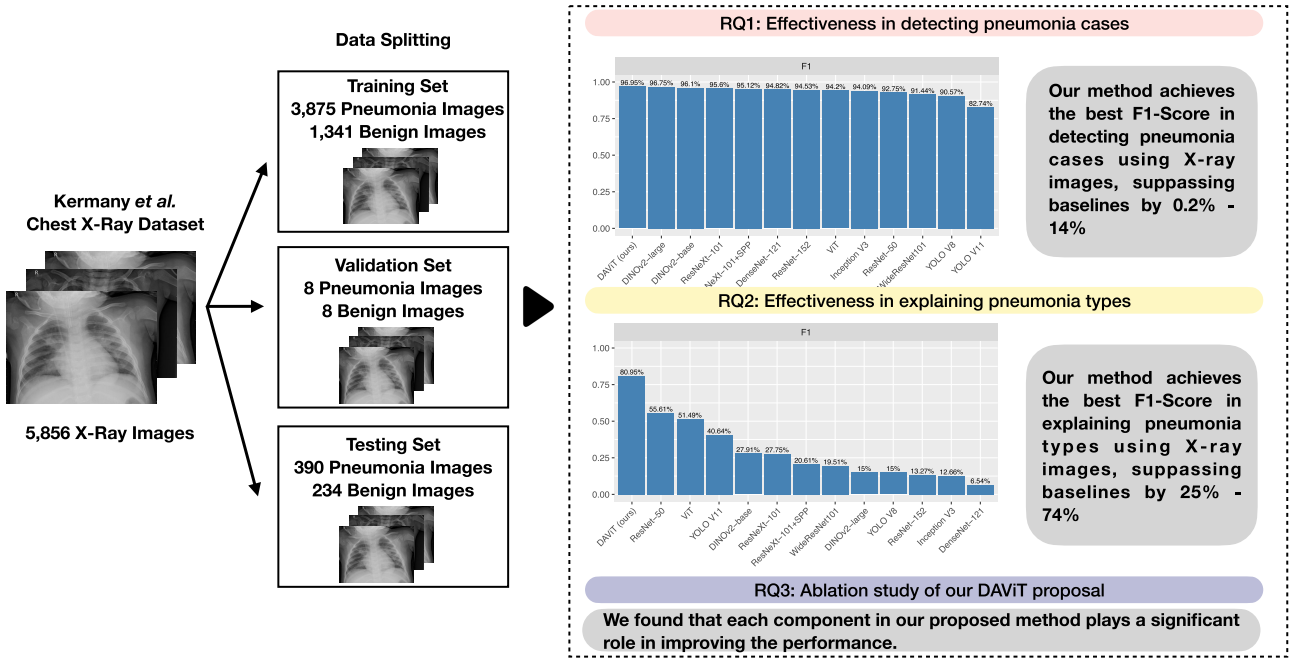

FIGURE 2. An overview of our experimental design, research questions, and result summary.

TABLE 1. (RQ1) The experimental results of our DAViT and the nine baseline comparisons for pneumonia detection using chest X-ray images. Results are presented in percentages. For inference time and FLOPs, (↓) lower = better; for all other metrics, (↑) higher = better. Abbreviations used: AUC (Area Under the Curve), ms (Milliseconds), and GigaFLOPs (Giga Floating Point Operations Per Second).

Method	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1	AUC	Inference Time Per Sample (ms)	FLOPs Per Sample (GigaFLOPs)
YOLO V8	87.98	88.89	92.31	80.77	90.57	86.54	0.12	9.44
YOLO V11	78.21	81.91	83.59	69.23	82.74	76.41	**0.11**	6.78
Inception V3	92.63	94.33	93.85	90.60	94.09	92.22	0.55	5.73
ResNet-50	91.19	95.39	90.26	92.74	92.75	91.50	0.64	4.11
ResNet-152	92.95	91.79	97.44	85.47	94.53	91.45	0.98	11.56
WideResNet101	89.58	94.04	88.97	90.60	91.44	89.79	1.22	22.80
ResNeXt-101	94.39	93.83	97.44	89.32	95.60	93.38	1.19	16.48
ResNeXt-101+SPP	93.91	95.36	94.87	92.31	95.12	93.59	0.74	16.47
DenseNet-121	93.43	93.52	96.15	88.89	94.82	92.52	0.91	**2.87**
ViT	92.47	90.74	97.95	83.33	94.20	90.64	1.06	16.87
DINOv2-base	95.03	94.32	97.95	90.17	96.10	94.06	0.89	27.78
DINOv2-large	95.83	94.39	**99.23**	90.17	96.75	94.70	1.82	98.42
DAViT (Ours)	**96.15**	**95.98**	97.95	**93.16**	**96.95**	**95.56**	12.77	100.85

refined by the CNN—facilitates effective feature learning across different levels of abstraction. Our ablation study in RQ3 further confirms that these architectural components contribute meaningfully to the model's effectiveness.

In terms of efficiency, YOLO V8 and V11 achieve the lowest inference latencies at 0.12 ms and 0.15 ms, respectively. Additionally, DenseNet-121 demonstrates the optimal GigaFLOPs of 2.87. These results indicate that while the model may not offer the highest accuracy, it achieves efficient performance with fewer operations, which can lead to faster inference and lower power consumption compared to more complex models. According to Table 1, larger models like DINOv2 and our DAViT approach achieve higher detection accuracy compared to CNN-based methods. However, there is a trade-off between detection accuracy and the inference time/computational power required.

B. (RQ2) HOW ACCURATE IS OUR DAViT FOR EXPLAINING PNEUMONIA TYPES USING CHEST X-RAY IMAGES?

Approach. To address this RQ, we evaluate the accuracy of DAViT for explaining pneumonia types using chest X-ray images. We compare DAViT with the same twelve baseline approaches as in RQ1. The goal of this task is to explain whether pneumonia is caused by a virus or bacteria, which we formulate as a binary classification problem. To this end, we use the same dataset as in RQ1, training the model on pneumonia images (excluding normal X-ray images), selecting the best model based on validation loss, and evaluating performance using the same testing set, which consists of 148 viral pneumonia images and 242 bacterial pneumonia images.

Results. Table 2 presents the experimental results for pneumonia type (i.e., viral or bacterial) explanations predicted by DAViT and the twelve baseline approaches, evaluated using

TABLE 2. (RQ2) The experimental results of our DAViT and the nine baseline comparisons for pneumonia type explanation using chest X-ray images. Results are presented in percentages. For inference time and FLOPs, (↓) lower = better; for all other metrics, (↑) higher = better. Abbreviations used: AUC (Area Under the Curve), ms (Milliseconds), and GigaFLOPs (Giga Floating Point Operations Per Second).

Method	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1	AUC	Inference Time Per Sample (ms)	FLOPs Per Sample (GigaFLOPs)
YOLO V8	65.13	**100.00**	8.11	**100.00**	15.00	54.05	**0.12**	9.44
YOLO V11	71.54	97.44	25.68	99.59	40.64	62.63	0.15	6.78
Inception V3	63.33	**100.00**	6.76	**100.00**	12.66	53.38	0.82	5.73
ResNet-50	76.67	**100.00**	38.51	**100.00**	55.61	69.26	0.98	4.11
ResNet-152	56.41	27.08	8.78	85.54	13.27	47.16	0.64	11.56
WideResNet101	66.15	**100.00**	10.81	**100.00**	19.51	55.41	1.22	22.80
ResNeXt-101	67.95	96.00	16.22	99.59	27.75	57.90	1.19	16.48
ResNeXt-101+SPP	66.41	**100.00**	11.49	**100.00**	20.61	55.74	1.06	16.47
DenseNet-121	63.33	**100.00**	3.38	**100.00**	6.54	51.69	0.91	**2.87**
ViT	74.87	96.30	35.14	99.17	51.49	67.15	1.06	16.87
DINOv2-base	68.21	**100.00**	16.22	**100.00**	27.91	58.11	0.89	27.78
DINOv2-large	65.13	**100.00**	8.11	**100.00**	15.00	54.05	1.82	98.42
DAViT (Ours)	**87.69**	98.08	**68.92**	99.17	**80.95**	**84.05**	12.77	100.85

six metrics: accuracy, precision, recall, specificity, F1-score, and AUC.

Our DAViT achieves an F1-score of 81%, outperforming state-of-the-art approaches by 25% to 74%, and an AUC of 84%, representing a 15% to 37% improvement over existing methods. Furthermore, our DAViT also achieves the best accuracy of 88%, recall of 69%, and comparable precision and specificity. The results demonstrate a substantial improvement over previous state-of-the-art pneumonia explanation methods, specifically the Inception V3 model utilized by Kermany et al. [3] and the DenseNet-121 model employed by Rajpurkar [4]. DAViT achieves a 68% and 74% increase in F1-score, respectively, compared to these models, along with a 31% and 32% improvement in AUC. These findings confirm that DAViT provides more effective pneumonia type explanations using chest X-ray images, demonstrating its ability to identify patterns that distinguish between viral and bacterial pneumonia with greater overall performance than existing deep learning-based approaches.

As with the results presented in RQ1, YOLO V8 and V11 achieve the lowest inference latencies, and DenseNet-121 demonstrates the optimal FLOPs. While our approach exhibits the highest inference latency and FLOPs, it excels in distinguishing between viral and bacterial pneumonia from X-ray images, achieving an F1-score of 81%, with a clear margin over all other baseline approaches. However, based on the current findings, future research should focus on investigating methods to compress the model to achieve acceptable performance while ensuring faster inference times and reduced computational power consumption.

C. (RQ3) WHAT ARE THE CONTRIBUTIONS OF THE COMPONENTS OF OUR DAViT?

Approach. To address this RQ, we investigate the contribution of each component within DAViT. Specifically, DAViT consists of three key components: a DINOv2 model pre-trained on the ImageNet dataset [6], a shallow Convolutional Neural Network (CNN), and domain adaptation using the CXR8 (ChestX-ray8) dataset collected by Wang et al. [23]. We use the same dataset as in RQ1, focusing on the pneumonia detection task to quantify the performance contribution of each component in DAViT. To this end, we extend our experiment to systematically evaluate the following five variants of pneumonia detection:

- *DINOv2-large Pre-Training + CNN + Domain Adaptation (DAViT)*: DINOv2-large pre-trained on ImageNet, CNN, and domain adaptation on the CXR8 dataset.
- *DINOv2-large Pre-Training + CNN*: DINOv2-large pre-trained on ImageNet and CNN, without domain adaptation.
- *DINOv2-large Pre-Training + Domain Adaptation*: DINOv2-large pre-trained on ImageNet and domain adaptation on the CXR8 dataset, without CNN.
- *DINOv2-large w/o Pre-Training + CNN + Domain Adaptation*: DINOv2-large without pre-training on ImageNet, but with CNN and domain adaptation.
- *DINOv2-large w/o Pre-Training*: DINOv2-large model without pre-training on ImageNet, CNN, and domain adaptation on the CXR8 dataset.

Results. Table 3 presents the ablation study to evaluate the contributions of the components of our DAViT.

The pre-training component of our DAViT is the most important component. Within DAViT, the pre-training component contributes 12% to both F1-score and AUC. When comparing DAViT with *DINOv2-large w/o Pre-Training + CNN + Domain Adaptation*, where the pre-training component is eliminated, we observe a performance decrease from 96.95% to 85.04% for F1-score, and from 95.56% to 84.02% for AUC. The CNN component within DAViT contributes 1% to both F1-score and AUC. When comparing DAViT with *DINOv2-large Pre-Training + Domain Adaptation*, where the CNN component is removed, we observe a performance decrease from 96.95% to 96.03% for F1-score, and from 95.56% to 94.66% for AUC. The domain adaptation component within DAViT contributes 0.2% to F1-score and 0.4% to AUC. When comparing DAViT with *DINOv2-large Pre-Training + CNN*, where the domain adaptation component is eliminated, we observe a performance decrease from 96.95% to 96.71% for F1-score, and from 95.56% to 95.13% for AUC.

Notably, the three components collectively contribute 21% to both F1-score and AUC. When comparing DAViT with

TABLE 3. (RQ3) The contribution of each component of DAViT for pneumonia detection using chest X-ray images. Results are presented in percentage. For all metrics, (↗) higher = better.

Method	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1	AUC
DINOv2-large Pre-Training + CNN + Domain Adaptation (ours)	**96.15**	**95.98**	**97.95**	**93.16**	**96.95**	**95.56**
DINOv2-large Pre-Training + CNN	95.83	95.5	**97.95**	92.31	96.71	95.13
DINOv2-large Pre-Training + Domain Adaptation	95.03	95.91	96.15	**93.16**	96.03	94.66
DINOv2-large w/o Pre-Training + CNN + Domain Adaptation	82.69	92.47	78.72	89.32	85.04	84.02
DINOv2-large w/o Pre-Training	73.24	86.32	67.95	82.05	76.04	75

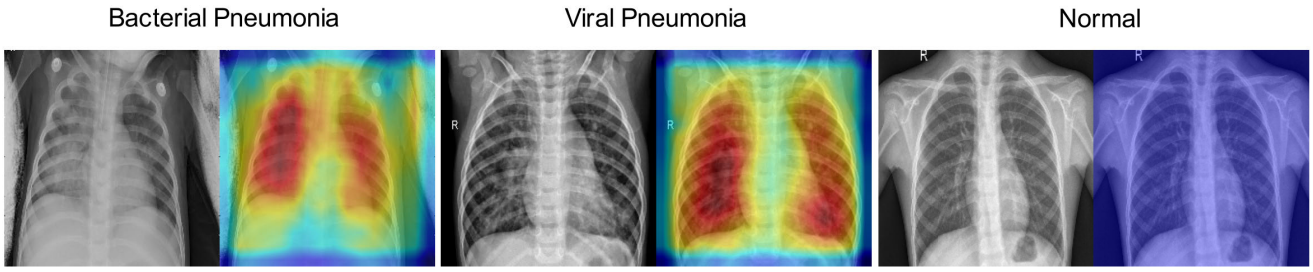

FIGURE 3. Heatmaps for bacterial pneumonia (left), viral pneumonia (middle), and normal case (right) when applying Grad-Cam to our DAViT approach. For bacterial and viral pneumonia, the heatmaps highlight regions in the lungs where our DAViT approach focused when correctly identifying pneumonia. For the normal case, the absence of highlights indicates the model correctly excluded pneumonia.

DINOv2-large w/o Pre-Training, where all three components are removed, we observe a performance decrease from 96.95% to 76.04% for F1-score, and from 95.56% to 75% for AUC. **These findings highlight that each component of DAViT plays a significant role in improving performance.**

VII. DISCUSSION

A. VISUAL EXPLANATIONS FOR MODEL DECISION-MAKING

In the previous section, we presented our experimental results and demonstrated the effectiveness of our DAViT approach in detecting pneumonia and explaining its types, outperforming state-of-the-art baselines. However, the model alone does not inherently provide localization for the specific regions in chest X-rays that the model focuses on during inference. This lack of visual explanation makes it challenging to understand the model's reasoning, particularly in medical contexts where transparency is critical for trust and adoption.

Providing localization for the regions of interest is crucial for clinical applications because it allows healthcare professionals to verify whether the model's focus aligns with known pathological features. To address this need, we leverage Grad-CAM, a visualization technique that generates heatmaps by identifying the areas in the input image that contribute most to the model's prediction [29]. Grad-CAM utilizes the gradients of the target output with respect to the model's feature maps to highlight important regions the model focuses on during inference. By applying this method, we can provide a detailed explanation of the model's reasoning process, showing the specific lung areas that influenced its decision. In what follows, we illustrate three distinct examples extracted from our testing dataset.

In the case of bacterial pneumonia (as shown on the left in Fig 3), the Grad-CAM heatmap highlights regions in the lung fields corresponding to areas of potential consolidation, a key feature of bacterial infections on chest X-rays [30]. These consolidations often appear as dense white opacities caused by fluid-filled alveoli and are typically localized to specific lobes. The model correctly predicted the label as bacterial pneumonia, and the heatmap focuses on central and peripheral lung regions where such abnormalities are expected. This alignment between the model's reasoning and clinical expectations suggests that the Grad-CAM visualization accurately identifies the areas of infection, providing interpretability for the model's decision-making process.

For the viral pneumonia case (as shown in the middle in Fig 3), the Grad-CAM heatmap highlights bilateral and diffuse regions in the lung fields, particularly around the perihilar areas. This pattern aligns with the characteristic radiological findings of viral pneumonia, which include diffuse interstitial infiltrates and ground-glass opacities [31]. The absence of intense, focal highlights supports the distinction from bacterial pneumonia, where localized consolidations are more common. The model correctly predicted the label as viral pneumonia, and the heatmap reasoning appears consistent with clinical findings. This result demonstrates that the Grad-CAM visualization successfully reflects the diffuse and bilateral nature of viral pneumonia.

In the normal case (as shown on the right in Fig 3), the model correctly predicted the absence of pneumonia. When Grad-CAM was applied with the target label set to 1 (pneumonia), the heatmap showed no significant activations or highlighted regions. This lack of focus indicates that the model found no evidence of pneumonia-related abnormalities, such as consolidations or diffuse infiltrates.

This Grad-CAM visualization supports the conclusion that the model confidently excludes pneumonia when the X-ray appears normal, enhancing its interpretability in clinical contexts.

B. CLINICAL DEPLOYMENT: CHALLENGES AND OPPORTUNITIES

While DAViT demonstrates strong performance, deploying it in diverse clinical settings presents several challenges.

First, variations in X-ray imaging protocols, equipment, and patient demographics across hospitals may affect model generalization, requiring careful domain adaptation. One may not directly apply the pre-trained DAViT presented in this paper; instead, further fine-tuning may be necessary. However, this process requires machine learning expertise, which most medical professionals lack, highlighting the need for low-code or no-code platforms to facilitate model adaptation in different clinical contexts.

Second, as shown in Table 1 and 2, while DAViT achieves the best performance in detecting pneumonia using chest X-ray images and is the only method to exceed an 80% F1-score for distinguishing viral and bacterial pneumonia cases, it also requires the highest inference time and FLOPs. This computational cost can be a burden, especially if additional fine-tuning is needed for deployment in a specific clinical setting, posing a challenge for real-world integration. Finally, regulatory and ethical considerations, including compliance with medical standards and patient privacy laws, must be addressed before clinical adoption.

To address these challenges, future research should explore the development of efficient fine-tuning strategies that reduce computational overhead while maintaining comparable performance. Additionally, designing intuitive, low-code AI platforms can empower medical practitioners to adapt models without extensive technical expertise. Further studies on model interpretability, fairness, and regulatory compliance will also be crucial for facilitating safe and effective deployment in real-world clinical environments.

VIII. LIMITATIONS

While our experiments are conducted on a carefully curated and widely used chest X-ray dataset [3], this does not guarantee the generalization of DAViT to other medical datasets or clinical settings. Differences such as imaging protocols and patient demographics across institutions may impact model performance. Thus, future research should explore automated pneumonia detection using chest X-ray images on other datasets. Another limitation of our approach is the relatively high computational cost, primarily due to using Vision Transformers. Future work will explore more efficient architectures and optimization techniques to reduce resource requirements.

IX. CONCLUSION

In this paper, we propose DAViT, a domain-adapted hybrid deep learning approach combining Vision Transformers (ViTs) and shallow Convolutional Neural Networks (CNNs) to improve pneumonia detection and type classification using chest X-ray images. Through an empirical evaluation of a real-world chest X-ray dataset comprising 5,856 images, we show that DAViT achieves (1) 96% of AUC for pneumonia detection and (2) 84% of AUC for pneumonia type explanation. These results demonstrate that DAViT outperforms existing state-of-the-art approaches for automated pneumonia detection and explanation tasks. Our ablation study confirms the effectiveness of each component in DAViT, including the ViT, CNN, and domain adaptation. Additionally, by applying Grad-CAM for visual explanations, we identify critical regions in the X-ray images that contribute to model predictions, providing interpretability for medical practitioners. Thus, we expect that DAViT can aid radiologists in making more accurate and interpretable automated pneumonia diagnoses.

ACKNOWLEDGMENT

The authors acknowledge the use of the OpenAI ChatGPT-4o-mini (free version) for grammar editing during the manuscript preparation. No content was generated by the AI beyond language refinement.

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] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-Rays with deep learning," 2017, *arXiv:1711.05225*.

[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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[11] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[13] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[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. Advancements Smart, Secure Intell. Comput. (ASSIC)*, Jan. 2024, pp. 1–5.

[15] M. H. Hosen, A. Saha, A. Uddin, K. Ashraf, and S. Nawar, "Enhancing pneumonia detection: CNN interpretability with LIME and SHAP," in *Proc. 6th Int. Conf. Electr. Eng. Inf. Commun. Technol. (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, D. Muhamediyeva, H. Primova, D. Sultanov, K. Tashev, and H. S. Jeon, "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: Multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung cancer, and tuberculosis using chest X-rays," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 13855–13880, Apr. 2023.
[19] A. Naeem, T. Anees, M. Fiza, R. A. Naqvi, and S.-W. Lee, "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, A. Naeem, R. A. Naqvi, and W.-K. Loh, "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, T. Anees, A. Naeem, R. A. Naqvi, and W.-K. Loh, "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, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (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 Conf. Comput. Vis. Pattern Recognit. (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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "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, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[30] S. B. A. Sattar, A. D. Nguyen, and S. Sharma, "Bacterial pneumonia," in *StatPearls*. Treasure Island, FI, 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.

MICHAEL FU received the Ph.D. degree from the Department of Software Systems and Cybersecurity, Monash University, in January 2025. He is currently a Lecturer (an Assistant Professor) with the Software Engineering Group, School of Computing and Information Systems, University of Melbourne. He has published nine papers in top-tier, peer-reviewed conferences in software engineering, including ICSE, FSE, ASE, and MSR, as well as esteemed journals, such as TSE, TOSEM, and EMSE. His research focuses on developing proactive security approaches for the DevSecOps lifecycle. He was invited to serve as a Program Committee (PC) Member for ICSE 2026, a PC Member for MSR 2024, a Junior PC Member, and the Web Co-Chair for MSR 2023, and a reviewer for papers submitted to TSE and TOSEM. His achievements include receiving the ACM SIGSOFT Distinguished Paper Award at ASE 2021 and the Distinguished Reviewer Award at MSR 2024.

CHAKKRIT TANTITHAMTHAVORN (Senior Member, IEEE) is currently a Senior Lecturer and the Director of Engagement and Impact, Faculty of Information Technology, Monash University, Australia. He is pioneering an emerging research area of explainable AI for software engineering, inventing many AI-based technologies to improve developers' productivity, and make software systems more reliable and more secure while being explainable to practitioners. He has made several major advances in explainable AI for software engineering and published the first online book on *Explainable AI for Software Engineering* (http://xai4se.github.io), attracting more than 13,000 page views from 83 countries worldwide and receiving positive responses from the SE community. His publications, books, and tutorials have informed many other studies and educated the SE community on the importance of explainability and its applications to software engineering. More about him at http://chakkrit.com

TRUNG LE (Member, IEEE) is currently a Lecturer with the Department of Data Science and AI, Monash University, Australia. He specializes in topics, such as optimal transport theory, machine learning, optimization, probabilistic inference, generative models, transfer learning, continual learning, adversarial/trustworthy machine learning, and cyber security. He has published over 100 papers in prestigious conferences and journals and secured over 1 million dollars in funding for research in areas, such as generative models, adversarial/trustworthy machine learning, transfer learning, and cyber-security.