

Article

AACNN-ViT: Adaptive Attention-Augmented Convolutional and Vision Transformer Fusion for Lung Cancer Detection

Mohammad Ishtiaque Rahman ^{1,*}  and Amrina Rahman ²¹ Thomas More University, Crestview Hills, KY 41017, USA² University of Dhaka, Dhaka 1000, Bangladesh; amrinarahman.work@gmail.com

* Correspondence: rahmanm@thomasmore.edu

Abstract

Lung cancer remains a leading cause of cancer-related mortality. Although reliable multiclass classification of lung lesions from CT imaging is essential for early diagnosis, it remains challenging due to subtle inter-class differences, limited sample sizes, and class imbalance. We propose an Adaptive Attention-Augmented Convolutional Neural Network with Vision Transformer (AACNN-ViT), a hybrid framework that integrates local convolutional representations with global transformer embeddings through an adaptive attention-based fusion module. The CNN branch captures fine-grained spatial patterns, the ViT branch encodes long-range contextual dependencies, and the adaptive fusion mechanism learns to weight cross-representation interactions to improve discriminability. To reduce the impact of imbalance, a hybrid objective that combines focal loss with categorical cross-entropy is incorporated during training. Experiments on the IQ-OTH/NCCD dataset (benign, malignant, and normal) show consistent performance progression in an ablation-style evaluation: CNN-only, ViT-only, CNN-ViT concatenation, and AACNN-ViT. The proposed AACNN-ViT achieved 96.97% accuracy on the validation set with macro-averaged precision/recall/F1 of 0.9588/0.9352/0.9458 and weighted F1 of 0.9693, substantially improving minority-class recognition (Benign recall 0.8333) compared with CNN-ViT (accuracy 89.09%, macro-F1 0.7680). One-vs.-rest ROC analysis further indicates strong separability across all classes (micro-average AUC 0.992). These results suggest that adaptive attention-based fusion offers a robust and clinically relevant approach for computer-aided lung cancer screening and decision support.

Keywords: lung cancer detection; attention mechanism; vision transformer; convolutional neural network; adaptive fusion; medical image analysis; CT imaging



Academic Editors: Lei Li and
Zehor Belkhatir

Received: 18 November 2025

Revised: 20 January 2026

Accepted: 24 January 2026

Published: 30 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Lung cancer remains one of the most fatal malignancies worldwide [1], with high morbidity and mortality rates largely attributed to late-stage detection. Early and accurate classification of lung abnormalities plays a vital role in improving the patient's prognosis and informing timely treatment strategies [2]. Computed tomography (CT) imaging has become the standard modality for lung cancer diagnosis due to its ability to capture high-resolution anatomical structures of the thoracic region [3]. However, interpreting these scans is inherently complex, involving subtle differences between benign, malignant, and normal tissues. This complexity, combined with the challenge of class imbalance and variability in scan quality, poses significant obstacles for automated classification systems.

Traditional deep learning models such as Convolutional Neural Networks (CNNs) have demonstrated considerable success in medical image analysis by learning spatially local features [4,5]. However, their limited receptive field restricts their ability to capture global context, which is crucial in differentiating between visually similar lesions [6]. Vision Transformers (ViTs), on the other hand, excel at modeling long-range dependencies and global representations through self-attention mechanisms [7]. Despite their advantages, ViTs are computationally intensive and often require large datasets [8], which may not always be feasible in medical imaging scenarios.

The integration of CNN and ViT offers a promising direction for capturing complementary local and global characteristics [9]. However, most existing fusion strategies rely on static concatenation or simple addition of features [10,11], which do not explicitly account for the relative relevance of CNN and ViT representations for each image. Furthermore, in the presence of class imbalance, which is common in medical datasets, standard loss functions may bias learning toward dominant classes and degrade performance on minority lesion types.

To address these challenges, this study explores a family of hybrid architectures that combine CNN and ViT features for multiclass lung cancer classification from CT images. Specifically, we design an Attention-Augmented Convolutional Neural Network with Vision Transformer (AACNN-ViT), which incorporates an adaptive attention mechanism to modulate the contributions of CNN and ViT feature embeddings on a per-image basis. In parallel, we construct non-adaptive CNN-ViT baselines that use fixed, concatenation-based fusion. A hybrid loss function combining focal loss and cross-entropy is employed across models to mitigate class imbalance and emphasize harder-to-classify cases.

The objective of this study is therefore not only to develop an attention-based hybrid CNN-ViT model, but also to systematically evaluate whether adaptive fusion offers measurable benefits over simpler fixed-fusion strategies under realistic data constraints. By comparing CNN, ViT, CNN-ViT, and AACNN-ViT variants on the same CT dataset and training protocol, we aim to identify architectures that provide a favorable balance between accuracy, class-wise performance, and practical deployability. Given the clinical importance of early and reliable detection [12], such comparative evidence can support the informed adoption of AI-based systems in lung cancer imaging workflows.

2. Attention in Image Processing: An Overview

Attention mechanisms have revolutionized image processing in deep learning by enabling models to selectively emphasize relevant features within complex visual data [13]. These mechanisms leverage linear transformations, similarity scores, and weighted aggregations to capture dependencies across image regions, offering a robust alternative to traditional convolutional operations [10].

Attention operates by assigning importance weights to different parts of an input, typically represented as a set of feature vectors. For an input matrix $X \in \mathbb{R}^{N \times D}$, where N denotes the number of spatial locations or tokens and D represents the feature dimension, attention computes a transformed output by highlighting relationships between elements [14]. The core principle involves modeling pairwise interactions through queries, keys, and values derived via learnable projections.

Given token embeddings $X \in \mathbb{R}^{n \times d}$, self-attention first forms queries, keys, and values via learnable linear projections:

$$Q = XW_q, \quad K = XW_k, \quad V = XW_v, \quad (1)$$

where $W_q, W_k, W_v \in \mathbb{R}^{d \times d_k}$ are trainable matrices and d_k is the key/query dimension. Scaled dot-product attention is then computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \quad (2)$$

Self-attention allows a model to weigh the importance of each input element relative to all others within the same sequence or image. Given an input X , the mechanism computes three projections as follows [15]:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D \times d}$ are weight matrices, and d is the dimension of the attention space, often equal to D or a reduced size. The attention scores are calculated as [16]:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right).$$

Here, $QK^\top \in \mathbb{R}^{N \times N}$ measures the similarity between query and key vectors, scaled by $\sqrt{d}$ to stabilize gradients. The output is then a weighted sum of values [17]:

$$\text{Self-Attention}(X) = AV,$$

where $A \in \mathbb{R}^{N \times N}$ and $V \in \mathbb{R}^{N \times d}$, yielding an output of shape $\mathbb{R}^{N \times d}$. This formulation captures global dependencies, enabling the model to focus on spatially distant but contextually relevant features.

Cross-attention extends this concept to relate two different input sets, such as features from distinct feature representations or network layers. For inputs $X_1 \in \mathbb{R}^{N_1 \times D_1}$ (queries) and $X_2 \in \mathbb{R}^{N_2 \times D_2}$ (keys and values), the projections are defined as [18]:

$$Q = X_1W_Q, \quad K = X_2W_K, \quad V = X_2W_V,$$

with $W_Q \in \mathbb{R}^{D_1 \times d}$, $W_K, W_V \in \mathbb{R}^{D_2 \times d}$. The attention scores are computed as [19]:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right),$$

where $A \in \mathbb{R}^{N_1 \times N_2}$ reflects the similarity between X_1 and X_2 . The output is:

$$\text{Cross-Attention}(X_1, X_2) = AV,$$

with shape $\mathbb{R}^{N_1 \times d}$, aggregating information from X_2 based on its relevance to X_1 . This mechanism excels at fusing heterogeneous representations, enhancing flexibility in image processing tasks.

A variant known as adaptive attention introduces a learnable modulation to refine the attention process. For inputs Q and K , a similarity metric (such as the cosine similarity), is computed [20]:

$$\text{sim}(Q, K) = \frac{\sum_i Q_i K_i}{\|Q\|_2 \|K\|_2}.$$

A trainable parameter λ adjusts the attention weights through an adaptive factor:

$$\text{Adaptive Factor} = (\text{sim}(Q, K) + \lambda)^\alpha,$$

where α controls the sharpness of the modulation. The weighted scores are then [21]:

$$\text{Weighted Scores} = \text{Adaptive Factor} \times \frac{QK^T}{\sqrt{d}},$$

followed by softmax normalization to obtain the attention weights. This adaptive scaling emphasizes contextually significant interactions, enhancing the model's ability to prioritize relevant features dynamically [22].

These attention mechanisms introduce non-linearity and global context beyond conventional convolutional models. Self-attention enables exhaustive pairwise interactions with $O(N^2)$ complexity, while cross-attention integrates multi-source data with $O(N_1N_2)$ [23]. The adaptive variant enhances robustness through a learnable modulation parameter λ . Together, they provide a mathematical foundation for advanced vision architectures by prioritizing salient regions and contextual relationships.

3. Proposed Adaptive Attention for CNN-ViT Feature Fusion

We propose an adaptive attention mechanism that is designed to improve the fusion of heterogeneous features from convolutional neural networks (CNNs) and Vision Transformers (ViTs). Unlike standard dot-product attention [24], which applies a fixed scaling factor to query–key interactions, our formulation dynamically modulates attention weights based on the contextual relevance between input representations [20]. This approach allows the model to emphasize semantically aligned regions across feature representations while suppressing weak or misleading associations. The adaptive formulation serves as the core of our AACNN-ViT framework, enabling richer and more flexible cross-branch feature integration across complementary representations.

Attention mechanisms compute similarity-based weights to prioritize informative elements in the input. Given a CNN-derived feature vector x_1 and ViT-derived vector x_2 , classical cross-attention projects these into query, key, and value spaces:

$$Q = W_Q x_1, \quad K = W_K x_2, \quad V = W_V x_2,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{in}} \times d}$ and d is the attention dimension. The output is computed via:

$$h_{\text{fused}} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V.$$

While effective, this formulation is statically scaled and fails to account for the varying semantic alignment of features across complementary representations. Our adaptive mechanism introduces a modulation factor that adjusts the attention weights based on input relevance.

3.1. Adaptive Attention Weighting Definition

To enhance sensitivity to feature interactions, we define the adaptive attention as:

$$A(x_1, x_2) = \text{softmax}\left((\rho(x_1, x_2) + \lambda)^\alpha \times \frac{QK^T}{\sqrt{d}}\right),$$

$$h_{\text{fused}} = A(x_1, x_2)V.$$

where $\rho(x_1, x_2) = \frac{\|x_1\|_2 \|x_2\|_2}{1 + |\langle x_1, x_2 \rangle|}$ measures feature relevance, $\lambda \in \mathbb{R}_{\geq 0}$ is a trainable parameter, and $\alpha > 0$ is a hyperparameter controlling scaling sharpness. The projections Q, K, V are as before, with all embeddings mapped to the same dimension $d = 256$. When x_1 and x_2 are

dissimilar (i.e., orthogonal), the relevance ρ is large, amplifying attention; when similar, the modulation is reduced, focusing on fine distinctions.

3.2. Properties of Adaptive Attention

The adaptive mechanism ensures continuity through smooth components like norms, projections, and softmax, promoting stable gradient flow. The relevance term further introduces scale sensitivity. If $x_1 \rightarrow cx_1$, then

$$\rho(cx_1, x_2) = \frac{c\|x_1\|_2\|x_2\|_2}{1 + c|\langle x_1, x_2 \rangle|},$$

showing that the attention weights adjust proportionally to feature magnitudes and alignment. Third, the trainable λ stabilizes the scaling and supports end-to-end optimization. Its gradient is:

$$\frac{\partial A}{\partial \lambda} = \alpha(\rho + \lambda)^{\alpha-1} \times \text{softmax}'\left((\rho + \lambda)^\alpha \times \frac{QK^T}{\sqrt{d}}\right) \times \frac{QK^T}{\sqrt{d}},$$

which remains well-behaved due to non-negativity and differentiability of all terms.

The parameter λ is a learnable scalar offset applied to the sample-level modulation term. Increasing λ increases the effective scaling of the attention logits, producing a sharper (more peaked) softmax distribution; decreasing λ yields a smoother distribution. In this formulation, λ controls the overall confidence/sharpness of attention for a given sample rather than changing the relative ranking of individual token pairs.

3.3. Relation to Classical Attention

Our formulation generalizes classical attention. When $\rho(x_1, x_2) + \lambda = 1$ and $\alpha = 1$, we recover:

$$A(x_1, x_2) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right),$$

equivalent to standard scaled dot-product attention. Thus, classical attention is a special case of our adaptive model, which expands its representational flexibility through relevance-aware modulation. In particular, orthogonal feature vectors yield $\rho \approx \|x_1\|_2\|x_2\|_2$, amplifying informative but spatially distant signals, potentially beneficial for fusing CNN and ViT outputs.

3.4. Enhanced Feature Fusion in AACNN-ViT

Within the AACNN-ViT architecture, CNN and ViT embeddings are projected to 256 dimensions and passed into the adaptive attention module. The output $h_{\text{fused}} \in \mathbb{R}^{256}$ is then used by a fully connected classification head, allowing the network to combine fine-grained (CNN) and contextual (ViT) features with image-dependent weights. Despite added complexity in modulation, the computational cost remains comparable to classical attention: $O(N_1 N_2 d)$, where N_1 and N_2 denote token lengths. As shown in our experiments, this fusion strategy improves lung cancer classification performance compared with using the CNN or ViT branch in isolation, and provides a principled mechanism for relevance-aware feature integration across complementary representations.

4. Methods

We developed the Attention-Augmented Convolutional Neural Network with Vision Transformer (AACNN-ViT), a feature-fusion framework that integrates convolutional features with precomputed Vision Transformer (ViT) embeddings using an adaptive fusion mechanism (Figure 1). In addition to the proposed AACNN-ViT, we imple-

mented three baselines: a stand-alone CNN, a stand-alone ViT feature classifier, and a simple concatenation-based CNN-ViT model. An ablation variant without adaptive gating (AACNN-ViT_noAdaptive) was also included to isolate the effect of learned fusion weights. The overall methodology consists of five components: dataset preparation and preprocessing, ViT feature extraction, CNN-based feature learning, feature fusion (fixed and adaptive), and model training and evaluation. The core AACNN-ViT fusion procedure is summarized in Algorithm 1.

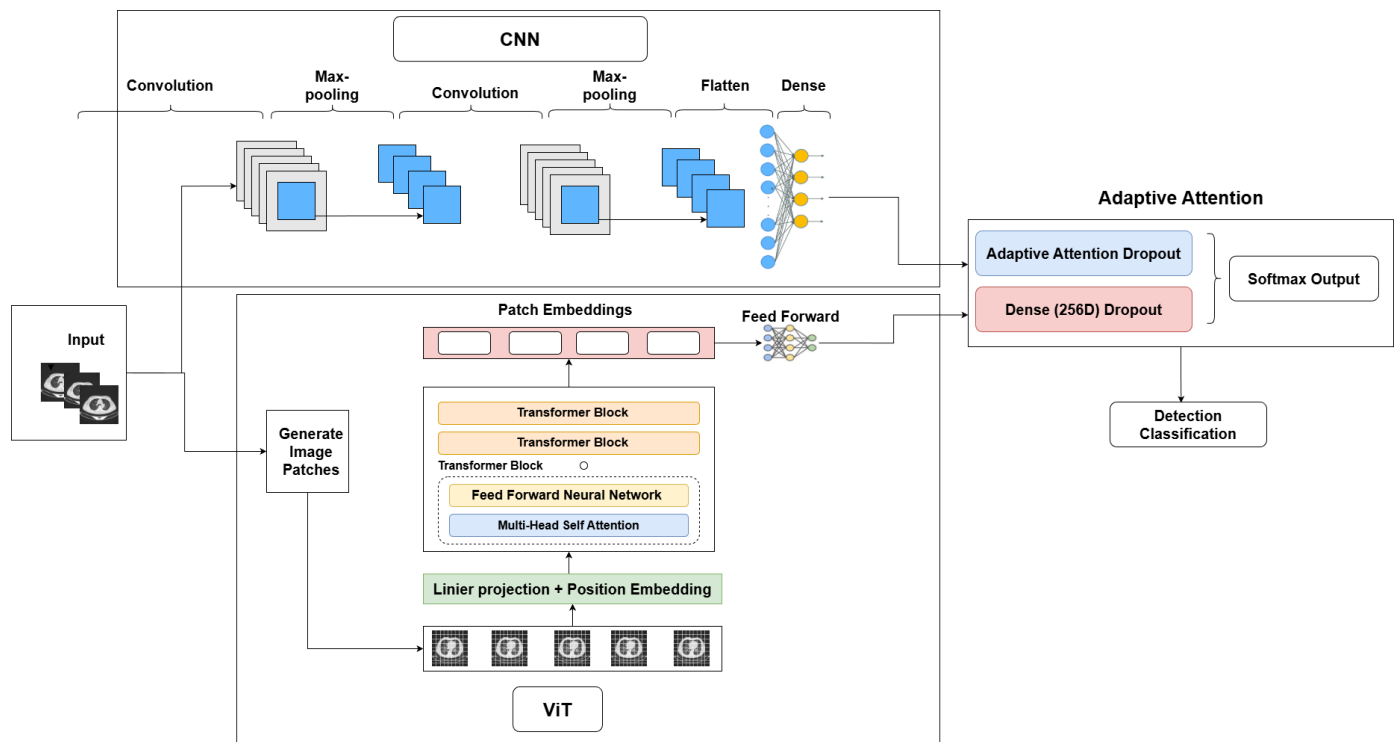


Figure 1. AACNN-ViT workflow for lung cancer classification, combining CNN and ViT features using adaptive attention.

4.1. Dataset and Preprocessing

The study utilized 1097 lung CT images from the Iraq–Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD) dataset [25]. The images are grouped into three diagnostic categories: Benign, Malignant, and Normal. To preserve class distribution, a stratified split was applied, allocating 70% of the samples (767 images) to the training set and the remaining 30% (330 images) to the validation set. All images were resized to a standardized resolution of $512 \times 512 \times 3$ pixels for processing by the convolutional branch. Class labels were encoded as one-hot vectors: $[1, 0, 0]$ for Benign, $[0, 1, 0]$ for Malignant, and $[0, 0, 1]$ for Normal.

To improve generalization and mitigate overfitting, data augmentation was performed using TensorFlow’s ImageDataGenerator. The augmentation pipeline included rescaling pixel intensities by $1/255$, random rotations up to 30° , width and height shifts up to 30%, shear and zoom transformations up to 30%, horizontal flips, brightness variations between $[0.8, 1.2]$, and channel shifts of up to 30 units. These transformations increased the variability of training samples while preserving diagnostic relevance.

Algorithm 1: Adaptive Gating CNN-ViT (AACNN-ViT) Training Procedure

Input: Training images $\mathcal{X} = \{x_i\}_{i=1}^{N_{\text{train}}}$,
precomputed ViT features $\mathcal{F} = \{f_i\}_{i=1}^{N_{\text{train}}}$,
labels $\mathcal{Y} = \{y_i\}_{i=1}^{N_{\text{train}}}$
Hyperparameters: batch size $B = 16$, max epochs $E = 20$, initial learning rate $\eta_0 = 5 \times 10^{-5}$,
weight decay $\lambda_w = 5 \times 10^{-6}$, focal loss parameters $\alpha = 0.25, \gamma = 2.0$,
hybrid loss weights $\beta_1 = 0.7, \beta_2 = 0.3$, early stopping patience $P = 8$, LR plateau window $P_{\text{LR}} = 3$
Output: Trained AACNN-ViT parameters Θ^*

- 1 Initialize CNN branch, ViT projection, gating network, and classifier parameters Θ randomly
- 2 Set learning rate $\eta \leftarrow \eta_0$
- 3 Load precomputed ViT features $\mathcal{F}$ into memory or on-disk cache
- 4 Set $L_{\text{best}} \leftarrow +\infty$, patience_ctr $\leftarrow 0$, plateau_ctr $\leftarrow 0$
- 5 **for** $e \leftarrow 1$ **to** E **do**
- 6 Shuffle $(\mathcal{X}, \mathcal{F}, \mathcal{Y})$
- 7 Partition into $M = \lceil N_{\text{train}}/B \rceil$ mini-batches
- 8 **for** $b \leftarrow 1$ **to** M **do**
- 9 Extract batch (X_b, F_b, Y_b) of size at most B
- 10 Apply data augmentation to X_b (rescaling, rotation, shifts, flips, etc.)
- 11 // Forward pass
- 12 $h_{\text{CNN},b} \leftarrow \phi_{\text{CNN}}(X_b)$ // CNN branch: conv + BN + pooling + GAP + dense to $\mathbb{R}^{256}$
- 13 $h_{\text{ViT},b} \leftarrow \phi_{\text{ViT}}(F_b)$ // ViT projection: MLP from $\mathbb{R}^{768}$ to $\mathbb{R}^{256}$
- 14 $z_b \leftarrow [h_{\text{CNN},b}, h_{\text{ViT},b}]$ // Concatenation in $\mathbb{R}^{512}$
- 15 // Adaptive gating
- 16 $g_b \leftarrow W_g \sigma(W_1 z_b + b_1) + b_g$ // Gating MLP outputs logits in $\mathbb{R}^2$
- 17 $[\alpha_b, \beta_b] \leftarrow \text{softmax}(g_b)$ // Per-sample weights, $\alpha_b + \beta_b = 1$
- 18 $h_{\text{fused},b} \leftarrow \alpha_b \odot h_{\text{CNN},b} + \beta_b \odot h_{\text{ViT},b}$ // Convex combination in $\mathbb{R}^{256}$
- 19 // Classifier head
- 20 $z'_b \leftarrow \text{Dropout}(h_{\text{fused},b})$
- 21 $z'_b \leftarrow \sigma(W_{\text{fuse}} z'_b + b_{\text{fuse}})$
- 22 $z'_b \leftarrow \text{Dropout}(z'_b)$
- 23 $\hat{Y}_b \leftarrow \text{softmax}(W_{\text{cls}} z'_b + b_{\text{cls}})$
- 24 // Hybrid loss
- 25 Compute categorical cross-entropy: $\text{CE}_b = -\frac{1}{B} \sum_{i=1}^B Y_{b,i} \log(\hat{Y}_{b,i})$
- 26 Compute focal loss: $p_{t,i} \leftarrow Y_{b,i} \hat{Y}_{b,i} + (1 - Y_{b,i})(1 - \hat{Y}_{b,i})$
- 27 FocalLoss $_b = -\frac{1}{B} \sum_{i=1}^B \alpha (1 - p_{t,i})^\gamma Y_{b,i} \log(\hat{Y}_{b,i})$
- 28 $L_b \leftarrow \beta_1 \cdot \text{FocalLoss}_b + \beta_2 \cdot \text{CE}_b$
- 29 // Parameter update (Adam with weight decay)
- 30 Compute gradients $\nabla_{\Theta} L_b$
- 31 Update Θ using Adam with learning rate η and weight decay λ_w
- 32 **end**
- 33 // Validation and scheduling
- 34 Compute validation loss L_{val} on the held-out 30% validation set (images + ViT features)
- 35 **if** $L_{\text{val}} < L_{\text{best}}$ **then**
- 36 $L_{\text{best}} \leftarrow L_{\text{val}}$
- 37 $\Theta^* \leftarrow \Theta$ // Store best parameters
- 38 patience_ctr $\leftarrow 0$
- 39 plateau_ctr $\leftarrow 0$
- 40 **end**
- 41 **else**
- 42 patience_ctr $\leftarrow$ patience_ctr + 1
- 43 plateau_ctr $\leftarrow$ plateau_ctr + 1
- 44 **end**
- 45 **if** plateau_ctr $\geq P_{\text{LR}}$ **then**
- 46 $\eta \leftarrow \eta/2$ // Reduce learning rate on plateau
- 47 plateau_ctr $\leftarrow 0$
- 48 **end**
- 49 **if** patience_ctr $\geq P$ **then**
- 50 // Early stopping
- 51 Break
- 52 **end**
- 53 **end**
- 54 **return** Θ^*

4.2. ViT Feature Extraction

We employed the pretrained google/vit-base-patch16-224-in21k model from Hugging Face to extract semantic features [26]. For this branch, each CT slice was resized to

$224 \times 224 \times 3$, normalized according to the ViT preprocessing pipeline, and passed through the network. The CLS token embeddings ($\mathbb{R}^{768}$) were extracted as global representations of each image [27].

To avoid redundant computation during training, these CLS features were precomputed and saved as NumPy arrays, yielding training features of shape (767,768) and validation features of shape (330,768). A shallow multilayer perceptron (MLP) operating directly on these ViT features served as the stand-alone ViT baseline.

4.3. CNN-Based Feature Learning

The CNN branch processes raw $512 \times 512 \times 3$ images to extract local spatial features [28]. It consists of three convolutional blocks with 32, 64, and 128 filters, respectively. Each block uses a 3×3 kernel with same padding and ReLU activation, followed by batch normalization and a 2×2 max-pooling operation. These operations progressively reduce the spatial resolution from 512×512 to 128×128 while increasing feature depth.

A global average pooling layer aggregates the final feature maps into a 128-dimensional vector, which is then projected to $\mathbb{R}^{256}$ via a dense layer with ReLU activation:

$$h_{\text{CNN}} = \phi_{\text{CNN}}(\text{GAP}(\text{ConvBlocks}(x))) \in \mathbb{R}^{256}.$$

A separate classification head applied directly to h_{CNN} is used as the stand-alone CNN baseline.

4.4. Feature Fusion via Fixed and Adaptive Attention

All fusion-based models operate on two 256-dimensional representations: the CNN projection $h_{\text{CNN}} \in \mathbb{R}^{256}$ and a projected ViT representation $h_{\text{ViT}} \in \mathbb{R}^{256}$. The ViT CLS embeddings $f_{\text{ViT}} \in \mathbb{R}^{768}$ are first mapped via a dense layer with GELU activation:

$$h_{\text{ViT}} = \phi_{\text{ViT}}(f_{\text{ViT}}) \in \mathbb{R}^{256}.$$

4.4.1. Simple Concatenation Fusion (CNN-ViT)

The CNN-ViT baseline fuses the two feature representations using simple concatenation:

$$z_{\text{concat}} = [h_{\text{CNN}}, h_{\text{ViT}}] \in \mathbb{R}^{512},$$

which is input to a fully connected head consisting of a dense layer with 256 units and ReLU activation, followed by dropout and a final softmax layer:

$$z = \sigma(W_{\text{fuse}}z_{\text{concat}} + b_{\text{fuse}}), \quad \hat{y} = \text{softmax}(W_{\text{cls}}z + b_{\text{cls}}).$$

This architecture serves as a strong fixed-fusion baseline.

4.4.2. Non-Adaptive Fusion (AACNN-ViT_noAdaptive)

AACNN-ViT_noAdaptive uses the same CNN and ViT branches and the same classification head as AACNN-ViT but without data-dependent gating. In this variant, the fusion layer reduces the concatenated vector to a fused representation through a dense transformation:

$$z_{\text{concat}} = [h_{\text{CNN}}, h_{\text{ViT}}], \quad h_{\text{fused}} = \sigma(W_{\text{red}}z_{\text{concat}} + b_{\text{red}}) \in \mathbb{R}^{256},$$

followed by dropout, a dense layer with 256 units, and the softmax classifier. This configuration controls for model capacity while removing adaptive weighting.

4.4.3. Adaptive Gating Fusion (AACNN-ViT)

The proposed AACNN-ViT introduces a lightweight adaptive gating mechanism that learns image-specific weights for the CNN and ViT embeddings (Algorithm 1). Rather than computing full token-wise self-attention, we model fusion as a learned convex combination of the two global feature vectors.

First, the embeddings are concatenated:

$$z = [h_{\text{CNN}}, h_{\text{ViT}}] \in \mathbb{R}^{512}.$$

A small gating network maps z to two logits:

$$g = W_g \sigma(W_1 z + b_1) + b_g \in \mathbb{R}^2,$$

where $W_1 \in \mathbb{R}^{64 \times 512}$ and $W_g \in \mathbb{R}^{2 \times 64}$. The logits are passed through a softmax to obtain non-negative weights that sum to one:

$$[\alpha, \beta] = \text{softmax}(g), \quad \alpha, \beta \geq 0, \quad \alpha + \beta = 1.$$

The fused representation is then given by:

$$h_{\text{fused}} = \alpha h_{\text{CNN}} + \beta h_{\text{ViT}} \in \mathbb{R}^{256}.$$

This fused vector is passed through a dropout layer (rate 0.4), a dense layer with 256 units and ReLU activation, an additional dropout layer (rate 0.3), and finally a softmax layer to produce three-class probabilities. Because the gating weights are learned from the joint representation z , AACNN-ViT can upweight CNN features when local texture patterns are more informative and upweight ViT features when global context is more salient.

4.5. Training and Optimization

Model training was performed using a custom generator that simultaneously produced batches of augmented images and their corresponding precomputed ViT embeddings (Algorithm 1). The same training protocol was applied to all models (CNN, ViT, CNN-ViT, AACNN-ViT_noAdaptive, and AACNN-ViT) to ensure a fair comparison.

A batch size of 16 was used throughout all experiments. Models were trained for up to 20 epochs using the Adam optimizer with an initial learning rate of 5×10^{-5} and a weight decay of 5×10^{-6} . To address class imbalance and stabilize convergence, we employed a hybrid loss function combining focal loss and categorical cross-entropy:

$$L(y, \hat{y}) = \beta_1 \times \text{FocalLoss}(y, \hat{y}) + \beta_2 \times \text{CE}(y, \hat{y}),$$

where the focal loss is defined as:

$$\text{FocalLoss} = -\frac{1}{B} \sum_{i=1}^B \alpha (1 - p_t)^\gamma y_i \log(\hat{y}_i), \quad p_t = y_i \hat{y}_i + (1 - \hat{y}_i)(1 - y_i),$$

and CE is the standard categorical cross-entropy:

$$\text{CE} = -\frac{1}{B} \sum_{i=1}^B y_i \log(\hat{y}_i).$$

The coefficients were set to $\alpha = 0.25$, $\gamma = 2.0$, $\beta_1 = 0.7$, and $\beta_2 = 0.3$, giving greater emphasis to hard-to-classify samples while maintaining stable optimization.

All models employed early stopping with a patience of 8 epochs, restoring the best-performing weights based on validation loss. Additionally, a learning rate reduction schedule was used, halving the learning rate when validation loss plateaued for 3 consecutive epochs. This adaptive training procedure promoted efficient convergence and reduced overfitting.

4.6. Evaluation Metrics

Model performance was assessed using six evaluation metrics: accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and loss [29]. These metrics collectively capture both overall predictive quality and class-wise discriminative ability.

Accuracy was computed as:

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

Precision and recall were defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.$$

The F1-score, representing the harmonic mean of precision and recall, was given by:

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

AUC values obtained from ROC curves quantified class separability across decision thresholds. Except for loss, all metrics were macro-averaged across the three classes—Benign, Malignant, and Normal—to ensure equal weighting regardless of class frequency.

Loss values reflected the hybrid objective function that combined focal loss with categorical cross-entropy, with parameters $\alpha = 0.25$, $\gamma = 2.0$, $\beta_1 = 0.7$, and $\beta_2 = 0.3$, as described above. This combined formulation allowed the models to focus more heavily on hard-to-classify samples while maintaining stable overall optimization.

4.7. Experimental Environment

Experiments were conducted on Google Colab Pro with an NVIDIA A100 GPU (40 GB HBM). The environment used Python 3.10, TensorFlow 2.16.1, Google's Vision Transformer (vit-base-patch16-224-in21k), NumPy 2.4.1, Matplotlib 3.10.8, and scikit-learn 1.8.0, leveraging GPU acceleration for all computations.

5. Results

We report an ablation study that progressively introduces (i) global ViT representations, (ii) simple CNN-ViT fusion via feature concatenation, and (iii) the proposed relevance-aware adaptive attention fusion. All models were evaluated on the same validation set ($n = 330$) spanning three classes (Benign, Malignant, Normal). Table 1 summarizes overall accuracy and macro-averaged precision, recall, and F1-score, while Table 2 provides per-class precision, recall, and F1-score. A consolidated visualization of the learning dynamics and class separability (accuracy, loss, and ROC curves) for all models is provided in Figure 2.

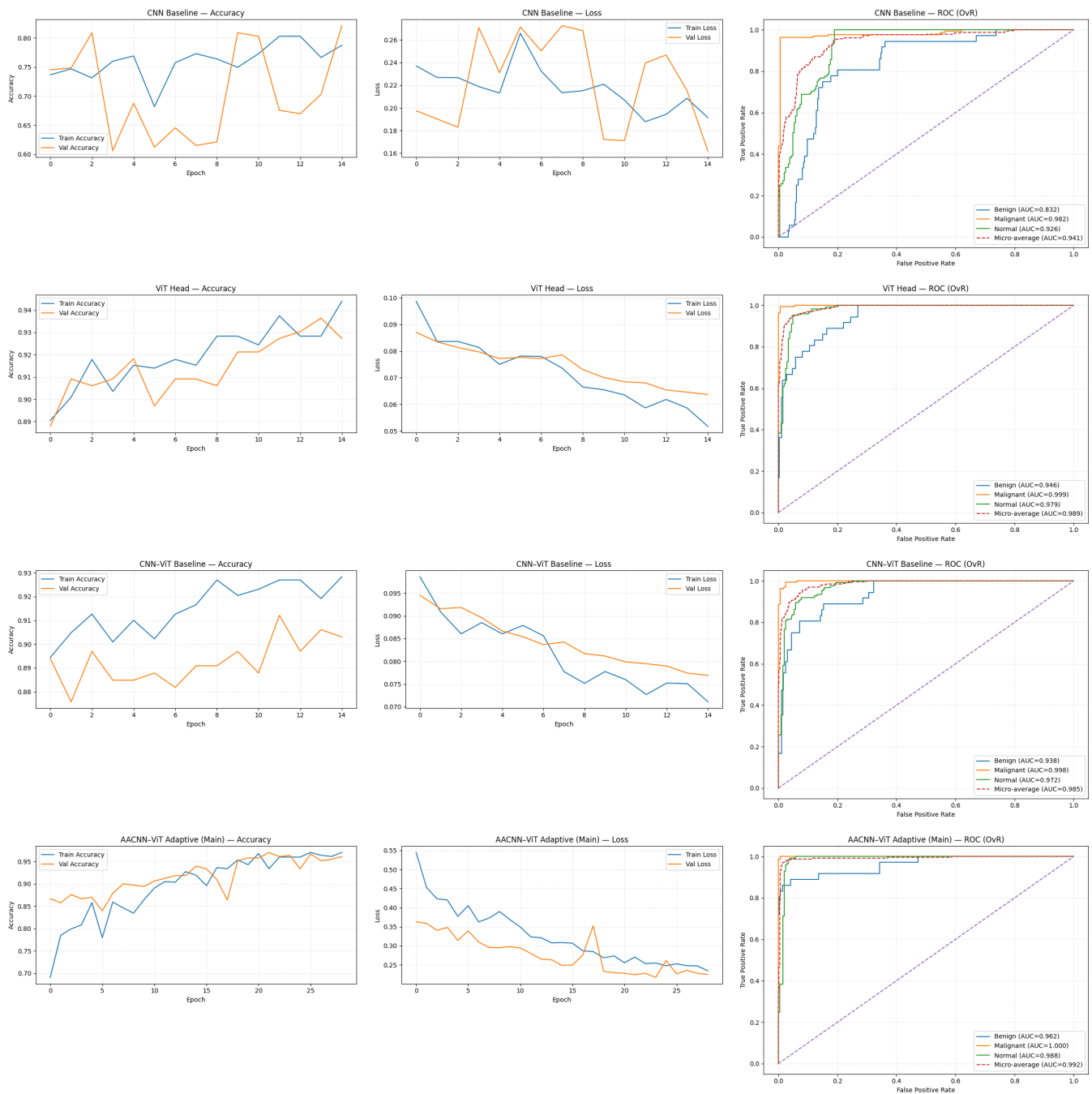


Figure 2. Training/validation accuracy and loss trajectories, and ROC curves (with AUC) for CNN, ViT, CNN-ViT, and AACNN-ViT on the validation set.

Table 1. Ablation over backbone and fusion strategy on lung cancer classification (validation set, $n = 330$). Metrics are macro-averaged across three classes.

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
CNN	0.6030	0.4923	0.4793	0.4215
ViT	0.8818	0.8294	0.7135	0.7248
CNN-ViT	0.8909	0.8440	0.7479	0.7680
AACNN-ViT (Adaptive)	0.9697	0.9588	0.9352	0.9458

Note: Bold indicates the best value in each metric column.

Table 2. Per-class performance (Precision, Recall, F1) on the validation set for each model.

Model	Class	Precision	Recall	F1-Score	No. of Img.
CNN	Benign	0.0000	0.0000	0.0000	36
	Malignant	0.9867	0.4379	0.6066	169
	Normal	0.4902	1.0000	0.6579	125
ViT Head	Benign	0.7273	0.2222	0.3404	36
	Malignant	0.9651	0.9822	0.9736	169
	Normal	0.7959	0.9360	0.8603	125
CNN-ViT	Benign	0.7500	0.3333	0.4615	36
	Malignant	0.9651	0.9822	0.9736	169
	Normal	0.8169	0.9280	0.8689	125
AACNN-ViT (Adaptive)	Benign	0.9375	0.8333	0.8824	36
	Malignant	1.0000	0.9882	0.9940	169
	Normal	0.9389	0.9840	0.9609	125

The CNN-only baseline, which relies exclusively on local convolutional features, achieved an accuracy of 60.30% and a macro F1-score of 0.4215 (Table 1). The per-class breakdown (Table 2) shows that this model failed to identify the Benign class (precision/recall/F1 all 0.0000), indicating a strong sensitivity to class imbalance and limited separability when only local cues are used. While the CNN produced perfect recall for Normal (1.0000), it did so with low precision (0.4902), reflecting substantial confusion between Normal and the other categories.

Replacing the CNN-only pipeline with a ViT-based head substantially improved performance, reaching 88.18% accuracy and a macro F1-score of 0.7248 (Table 1). This gain highlights the value of global contextual representations captured by pretrained ViT embeddings. Class-wise results show near-ceiling performance for Malignant (F1 = 0.9736) and strong performance for Normal (F1 = 0.8603), but Benign remains the most challenging class (F1 = 0.3404), consistent with its smaller support in the validation set.

The CNN-ViT model, which concatenates CNN features and ViT embeddings prior to classification, yields a modest additional improvement over the ViT-only head, achieving 89.09% accuracy and a macro F1-score of 0.7680 (Table 1). The per-class metrics indicate that adding local CNN information improves Benign detection (F1 increases from 0.3404 to 0.4615) while maintaining strong Malignant and Normal performance (Table 2). However, the improvement remains limited, suggesting that static concatenation does not fully exploit cross-representation dependencies or resolve residual minority-class confusion.

The proposed AACNN-ViT (Adaptive) achieves the best performance across all metrics, with 96.97% accuracy and a macro F1-score of 0.9458 (Table 1). Unlike static concatenation, the adaptive fusion yields consistently high and more balanced per-class outcomes, improving Benign F1 to 0.8824 while also strengthening Malignant and Normal performance (F1 = 0.9940 and 0.9609, respectively; Table 2). These results indicate that relevance-aware fusion between CNN and ViT representations is substantially more effective than either representation alone or naive feature concatenation.

To further characterize error modes, Figure 3 reports the confusion matrix for AACNN-ViT (Adaptive). The model correctly classifies 30/36 Benign cases, 167/169 Malignant cases, and 123/125 Normal cases, with the remaining errors concentrated primarily between Benign and Normal (6 Benign→Normal and 2 Normal→Benign). Malignant is rarely confused with other categories (2 Malignant→Normal and 0 Malignant→Benign), indicating strong separability for clinically critical malignant findings under the proposed fusion mechanism.

AACNN-ViT (Adaptive) Confusion Matrix (Validation Set)

True label	Benign	30	0	6
	Malignant	0	167	2
	Normal	2	0	123
		Benign	Malignant	Normal
		Predicted label		

Figure 3. Confusion matrix for AACNN-ViT (Adaptive) on the validation set ($n = 330$). Rows denote true labels and columns denote predicted labels. The model shows near-perfect discrimination for Malignant (167/169 correct), with remaining errors concentrated between Benign and Normal (6 Benign→Normal, 2 Normal→Benign).

6. Discussion

This study evaluated an adaptive fusion framework (AACNN-ViT) that integrates CNN-derived local representations with ViT-derived global embeddings for three-class lung cancer image classification (Benign, Malignant, Normal). The updated results demonstrate a clear progression in performance across an ablation-style comparison: CNN-only, ViT-only, CNN + ViT concatenation, and the proposed adaptive fusion. In particular, AACNN-ViT achieved the strongest overall validation performance, reaching 96.97% accuracy with a macro-F1 of 0.9458 and weighted-F1 of 0.9693. This represents a substantial improvement over the best baseline (CNN-ViT, 89.09% accuracy; macro-F1 0.7680) and indicates that the adaptive fusion contributes meaningfully beyond simply combining representations.

The CNN-only baseline underperformed primarily due to its failure to recover the minority Benign class. The confusion matrix shows that the CNN predicted virtually no Benign samples correctly (0/36), instead mapping Benign instances almost entirely to the Normal class. This collapse is also reflected in the Benign precision/recall/F1 of 0.0, while Malignant recall remained moderate and Normal recall was perfect. This pattern is consistent with a model dominated by majority-class decision boundaries, where local texture cues alone are insufficient to separate minority benign patterns from normal appearance in a limited dataset.

Replacing the CNN classifier with a ViT-feature head produced an immediate and substantial improvement in overall performance (88.18% accuracy; macro-F1 0.7248). The ViT head achieved strong discrimination for Malignant and Normal (Malignant F1 0.9736; Normal F1 0.8603), suggesting that global contextual encoding is critical for separating malignant morphology and normal structure. However, minority sensitivity remained limited:

Benign recall was only 0.2222 (8/36), indicating that global embeddings alone still do not fully capture the fine-grained characteristics needed to reliably detect benign findings.

The CNN-ViT concatenation baseline further improved performance (89.09% accuracy; macro-F1 0.7680), confirming that combining local and global information is beneficial. Nevertheless, the improvement in Benign detection was modest (Benign recall 0.3333), and most Benign errors still shifted to Normal. This suggests that static fusion (feature concatenation) does not adequately control how the two representation spaces interact; the classifier may continue to prioritize majority-aligned cues, especially when Benign samples are scarce.

AACNN-ViT delivered the most decisive improvement, particularly for the minority class. The model achieved Benign precision of 0.9375 and recall of 0.8333 (30/36), while also maintaining near-ceiling performance for Malignant (F1 0.9940) and Normal (F1 0.9609). The accompanying confusion matrix shows that residual errors are limited and clinically plausible: the remaining Benign misclassifications occur primarily as Normal (6 cases), while Malignant errors are rare (2 cases as Normal), and Normal is occasionally confused with Benign (2 cases). This distribution indicates that the adaptive fusion mechanism improves minority separation without degrading discrimination on majority classes, yielding a balanced classifier rather than a majority-optimized one.

These findings support the central claim that adaptive fusion is not a cosmetic modification but a functional component that materially changes the decision behavior of the hybrid model. Empirically, the gains concentrate where naive fusion fails most: minority recognition and error reduction in the Benign-vs.-Normal boundary. In practice, this is a meaningful improvement because benign lesions are often the hardest to detect reliably under class imbalance, and false reassurance (Benign→Normal) is a clinically undesirable error mode.

Figure-level diagnostics are also consistent with the quantitative improvements. Across models, the accuracy and loss trajectories show progressively more stable convergence as global context is introduced and as fusion becomes more targeted. The ROC curves similarly indicate strong separability across classes for the adaptive fusion model, aligning with its near-ceiling per-class F1 scores. Collectively, these results suggest that AACNN-ViT learns a representation that is both discriminative and robust to the skewed class distribution.

Explainability is a critical requirement for clinical adoption because deep neural networks, despite strong predictive performance, do not inherently provide transparent or human-interpretable decision rules in the same way as “shallow” pipelines (e.g., handcrafted radiomic features with classical machine learning). Prior work has highlighted that handcrafted imaging biomarkers can offer more direct interpretability at the feature level (e.g., texture, shape, intensity statistics), whereas deep learned representations are often harder to map to clinically meaningful descriptors without dedicated explanation techniques [30,31]. Although our study primarily emphasizes classification performance and efficiency rather than explicit explainability mechanisms, AACNN-ViT is compatible with established eXplainable AI (XAI) approaches that can be incorporated in future clinical-facing deployments (e.g., saliency visualization for the convolutional branch or attribution-based explanations for transformer representations), and the adaptive fusion behavior itself provides an intuitive indicator of how local and global evidence may contribute to predictions, supporting more transparent model auditing and safer decision-support use cases.

Several limitations remain. First, the minority class count (Benign) is small, and performance estimates can therefore be optimistic if evaluated on a single split. Reporting variability via repeated stratified splits or bootstrap confidence intervals, and adding

statistical tests between the best baseline and AACNN-ViT, would strengthen claims of robustness. Second, although precomputed ViT embeddings improve efficiency, they limit end-to-end adaptation of the transformer representation to the target domain; fine-tuning (when computationally feasible) may yield further gains but must be carefully controlled to avoid overfitting. Finally, the remaining Benign→Normal errors indicate that some benign patterns remain close to normal appearance; future work can address this through targeted sampling (oversampling/augmentation for Benign), cost-sensitive training, and explicit calibration analysis to reduce high-confidence minority misses.

7. Conclusions

We proposed AACNN-ViT, a hybrid architecture that integrates CNN and Vision Transformer features through adaptive attention for lung cancer classification from CT images. The model effectively balances local and global representations and addresses class imbalance using a hybrid loss function. Experimental results demonstrated superior performance over baseline models in terms of accuracy, precision, recall, and F1-score. AACNN-ViT shows strong potential for aiding clinical diagnosis, offering a scalable and robust approach for medical image analysis. Future work will focus on end-to-end optimization and explainability to improve clinical integration and trust.

Author Contributions: Conceptualization, M.I.R.; Methodology, M.I.R.; Validation, M.I.R.; Formal analysis, M.I.R. and A.R.; Investigation, A.R.; Resources, M.I.R.; Writing—original draft, M.I.R. and A.R.; Visualization, A.R.; Supervision, M.I.R.; Project administration, M.I.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the data used were obtained from the public databases.

Informed Consent Statement: Patient consent was waived due to the data used were obtained from the public databases.

Data Availability Statement: The data presented in this study are openly available in Mendeley at <https://data.mendeley.com/datasets/bhmdr45bh2/3> (accessed on 23 January 2025). All preprocessing scripts and model training code are available at <https://github.com/PublicDataSage/AACNN-ViT.git> (accessed on 23 January 2025).

Acknowledgments: During the preparation of this work the authors used ChatGPT 5.1 to improve the quality of writing. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tao, M.H. Epidemiology of lung cancer. In *Lung Cancer and Imaging*; IOP Publishing: Bristol, UK, 2019; pp. 4–14–15.
2. Inage, T.; Nakajima, T.; Yoshino, I.; Yasufuku, K. Early lung cancer detection. *Clin. Chest Med.* **2018**, *39*, 45–55. [[CrossRef](#)] [[PubMed](#)]
3. Makaju, S.; Prasad, P.; Alsadoon, A.; Singh, A.; Elchouemi, A. Lung cancer detection using CT scan images. *Procedia Comput. Sci.* **2018**, *125*, 107–114. [[CrossRef](#)]
4. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [[CrossRef](#)] [[PubMed](#)]
5. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical image analysis using convolutional neural networks: A review. *J. Med. Syst.* **2018**, *42*, 226. [[CrossRef](#)]
6. Rangel, G.; Cuevas-Tello, J.C.; Nunez-Varela, J.; Puente, C.; Silva-Trujillo, A.G. A survey on convolutional neural networks and their performance limitations in image recognition tasks. *J. Sens.* **2024**, *2024*, 2797320. [[CrossRef](#)]

7. Poonia, R.C.; Al-Alshaikh, H.A. Ensemble approach of transfer learning and vision transformer leveraging explainable AI for disease diagnosis: An advancement towards smart healthcare 5.0. *Comput. Biol. Med.* **2024**, *179*, 108874. [[CrossRef](#)] [[PubMed](#)]
8. Alabdulmohsin, I.M.; Zhai, X.; Kolesnikov, A.; Beyer, L. Getting vit in shape: Scaling laws for compute-optimal model design. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 16406–16425.
9. Jiang, Z.; Dong, Z.; Wang, L.; Jiang, W. Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model. *Comput. Intell. Neurosci.* **2021**, *2021*, 7529893. [[CrossRef](#)] [[PubMed](#)]
10. Li, X.; Li, M.; Yan, P.; Li, G.; Jiang, Y.; Luo, H.; Yin, S. Deep learning attention mechanism in medical image analysis: Basics and beyonds. *Int. J. Netw. Dyn. Intell.* **2023**, *2*, 93–116. [[CrossRef](#)]
11. Rahman, M.I. Fusion of Vision Transformer and Convolutional Neural Network for Explainable and Efficient Histopathological Image Classification in Cyber-Physical Healthcare Systems. *J. Transform. Technol. Sustain. Dev.* **2025**, *9*, 8. [[CrossRef](#)]
12. Magrabi, F.; Ammenwerth, E.; McNair, J.B.; De Keizer, N.F.; Hyppönen, H.; Nykänen, P.; Rigby, M.; Scott, P.J.; Vehko, T.; Wong, Z.S.Y.; et al. Artificial intelligence in clinical decision support: Challenges for evaluating AI and practical implications. *Yearb. Med. Inform.* **2019**, *28*, 128–134. [[CrossRef](#)] [[PubMed](#)]
13. Haut, J.M.; Paoletti, M.E.; Plaza, J.; Plaza, A.; Li, J. Visual attention-driven hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8065–8080. [[CrossRef](#)]
14. Nikzad, N.; Liao, Y.; Gao, Y.; Zhou, J. SATA: Spatial Autocorrelation Token Analysis for Enhancing the Robustness of Vision Transformers. *arXiv* **2024**, arXiv:2409.19850. [[CrossRef](#)]
15. Yang, M.; Ma, M.Q.; Li, D.; Tsai, Y.H.H.; Salakhutdinov, R. Complex transformer: A framework for modeling complex-valued sequence. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 4232–4236.
16. Qi, X.; Wang, T.; Liu, J. Comparison of support vector machine and softmax classifiers in computer vision. In Proceedings of the 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 8–10 December 2017; IEEE: New York, NY, USA, 2017; pp. 151–155.
17. Guo, M.H.; Liu, Z.N.; Mu, T.J.; Hu, S.M. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5436–5447. [[CrossRef](#)]
18. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2019; pp. 603–612.
19. Hao, Y.; Dong, L.; Wei, F.; Xu, K. Self-attention attribution: Interpreting information interactions inside transformer. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 12963–12971. [[CrossRef](#)]
20. Ma, X.; Hu, K.; Sun, X.; Chen, S. Adaptive Attention Module for Image Recognition Systems in Autonomous Driving. *Int. J. Intell. Syst.* **2024**, *2024*, 3934270. [[CrossRef](#)]
21. Okamoto, T.; Toda, T.; Shiga, Y.; Kawai, H. Transformer-based text-to-speech with weighted forced attention. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: New York, NY, USA, 2020; pp. 6729–6733.
22. Lin, H.; Bai, R.; Jia, W.; Yang, X.; You, Y. Preserving dynamic attention for long-term spatial-temporal prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery, New York, NY, USA, 2020; pp. 36–46.
23. Yu, T.; Khalitov, R.; Cheng, L.; Yang, Z. Paramixer: Parameterizing mixing links in sparse factors works better than dot-product self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE: New York, NY, USA, 2022; pp. 691–700.
24. Namazifar, M.; Hazarika, D.; Hakkani-Tür, D. Role of bias terms in dot-product attention. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
25. Alyasriy, H.; Al-Huseiny, M. The IQ-OTH/NCCD Lung Cancer Dataset. 2023. Available online: <https://data.mendeley.com/datasets/bhmdr45bh2/3> (accessed on 3 January 2025).
26. Google Research. ViT Base Patch16 224 (ImageNet-21k). 2022. Available online: <https://huggingface.co/google/vit-base-patch16-224-in21k> (accessed on 24 March 2024).
27. Zou, Y.; Yi, S.; Li, Y.; Li, R. A Closer Look at the CLS Token for Cross-Domain Few-Shot Learning. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 85523–85545.
28. Zheng, Y.; Huang, J.; Chen, T.; Ou, Y.; Zhou, W. CNN classification based on global and local features. In *Proceedings of the Real-Time Image Processing and Deep Learning 2019*; SPIE: Bellingham, WA, USA, 2019; Volume 10996, pp. 96–108.
29. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. [[CrossRef](#)] [[PubMed](#)]

30. Rundo, L.; Militello, C. Image biomarkers and explainable AI: Handcrafted features versus deep learned features. *Eur. Radiol. Exp.* **2024**, *8*, 130. [[CrossRef](#)] [[PubMed](#)]
31. Abrantes, J.; Rouzrokh, P. Explaining explainability: The role of XAI in medical imaging. *Eur. J. Radiol.* **2024**, *173*, 111389. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.