PulmoNet: A Hybrid CNN-Vision Transformer Architecture for Enhanced Lung Nodule Classification in CT Imaging

Venkatesh M R
Presidency School of Computer Science and Engineering, Presidency University, Bengaluru, India
VENKATESH.20233CSE0003@presidencyuniversity.in

Hasan Hussain Shahul Hameed
Presidency School of Computer Science and Engineering, Presidency University, Bengaluru, India
hasan.hussain@presidencyuniversity.in (corresponding author)

Received: 30 November 2025 | Revised: 25 December 2025 and 14 January 2026 | Accepted: 17 January 2026

 DOI: https://doi.org/10.48084/etasr.16612

ABSTRACT

Lung cancer remains the leading cause of cancer-related mortality, with early detection by CT screening being critical for patient survival. Current deep learning approaches face significant limitations: Convolutional Neural Networks (CNNs) extract local texture patterns but cannot capture global spatial relationships, while Vision Transformers (ViTs) model long-range dependencies but struggle with fine-grained feature extraction. Existing hybrid architectures use static fusion strategies that fail to adapt to diverse nodule characteristics. This paper presents PulmoNet, a novel hybrid framework that integrates a modified ResNet-50 with Vision Transformers through an adaptive cross-attention fusion mechanism that dynamically adjusts branch contributions based on individual nodule morphology. The framework processes CT volumes through parallel pipelines where CNNs extract multi-scale local patterns and transformers capture long-range spatial dependencies. Evaluated on the LUNA16 and LIDC-IDRI datasets using 5-fold cross-validation, PulmoNet achieves 94.7% accuracy and 0.982 AUC-ROC, outperforming state-of-the-art baselines by 3.5-5.4%. Cross-dataset evaluation demonstrates robust generalization across nodule sizes, types, and locations. PulmoNet demonstrates clinical viability with 93.8% sensitivity at 95% specificity and 143 ms inference time, establishing potential for real-time lung cancer screening programs.

Keywords-lung nodule classification; hybrid deep learning; Vision Transformer (ViT); adaptive fusion; medical image analysis; computer-aided diagnosis

I. INTRODUCTION

Lung cancer is one of the leading causes of death, accounting for approximately 1.8 million deaths annually worldwide [1]. Patient survival rates are highly dependent on the detection of the disease at different stages, ranging from 56% for localized tumors to merely 5% for metastatic cases over five years. Low-dose computed tomography screening can reduce mortality by 20-30%, making early detection critical for the treatment of lung cancer [2]. However, CT scan interpretation presents significant challenges, as each scan contains hundreds of axial slices requiring careful review to identify and characterize pulmonary nodules ranging from 3 to 30 mm in diameter. This process generally takes about 15-30 minutes per scan and reflects a substantial interobserver variability, with malignancy assessment agreement being frequently below 65% between radiologists [3].

The main complicating factor for accurate nodular classification is the visual similarity between benign and malignant nodules, particularly in small lesion dimensions. Convolutional Neural Networks (CNNs) have been very successful in learning hierarchical features directly from imaging data while capturing local texture patterns, edge information, and fine-grained spatial details that characterize nodule morphology [4]. However, CNNs process visual information through localized receptive fields, which restricts their ability to model long-range spatial dependencies within the lung parenchyma. Vision Transformers (ViTs) represent an alternative approach, using self-attention mechanisms that capture global spatial relationships by treating image patches as a sequence of tokens [5]. Although transformers overcome the limitation of CNNs in global context modeling, they are limited by their inability to extract local fine-grained features and require larger training datasets. The complementary nature of these two architectures indicates that a hybrid design that combines CNN-based local feature extraction with transformer-based global context modeling may improve lung nodule classification performance [6].

Deep learning applications in the medical imaging field were considerably expanded when CNNs demonstrated

outstanding performance in computer vision tasks. In [7], a categorization between feature-based machine learning and image-based methods, which process raw imaging data directly, showed that CNNs are a dominant approach for analyzing medical images. Their key advantages lie in learning directly from image data without manual feature engineering. Even after these advantages, standard CNNs face limitations due to localized convolutional operations that fail to capture long-range dependencies, which are essential to identify subtle indicators of malignancy.

In [8], attention mechanisms were investigated to overcome such constraints. A CNN architecture based on dual attention mechanisms combined channel and spatial attention to highlight informative features, achieving strong accuracy on benchmark datasets. In [9], a hybrid CNN with attention mechanisms for prognostic prediction across multiple lung diseases reported high accuracy on both CT and X-ray datasets. These attention-enhanced architectures still process information through local receptive fields, limiting their ability to capture global contextual relationships within thoracic CT images.

ViTs introduced a different approach, using self-attention mechanisms on image patch sequences. In [10], ViT-based models achieved 97.83% accuracy on chest X-ray classification, exceeding traditional CNN baselines. However, this work focused on X-ray imaging rather than CT scans, while using the transformer architecture alone does not consider the fine-grained local texture information crucial for distinguishing nodule characteristics. In [11], a review on deep learning for lung cancer diagnosis with CT imaging examined 80 studies reporting detection rates above 95% and classification accuracy of 99%. This review noted that hybrid architectures combining CNNs with transformers show promise, but less than 15% of the studies validated models across diverse populations or multiple clinical centers. In addition, challenges in dataset availability, annotation consistency, and generalizability remain open issues [12].

Despite the existing literature advancing the area of lung nodule classification, several technical aspects have not yet been fully explored [13]. In this respect, the proposed framework, PulmoNet, aimed to address the following objectives:

- Develop an automated pipeline that integrates preprocessing, U-Net-based lung segmentation, and a hybrid CNN-ViT classifier for end-to-end lung nodule classification.
- Combine ResNet-50 convolutional layers with ViT blocks to extract both local texture features and global spatial relationships from CT images within a unified architecture.
- Evaluate the hybrid model on LUNA16 and LIDC-IDRI benchmark datasets against CNN-only and ViT-only architectures using accuracy, precision, recall, F1-score, and AUC-ROC metrics.
- Perform systematic ablation studies to quantify the individual contribution of the CNN and ViT components to overall classification performance.

Figure 1 illustrates the PulmoNet framework, which processes CT scan images through preprocessing and U-Net segmentation, followed by parallel feature extraction using the ResNet-50 and ViT models. The extracted features are fused and classified to determine whether the lung nodules are benign or malignant.

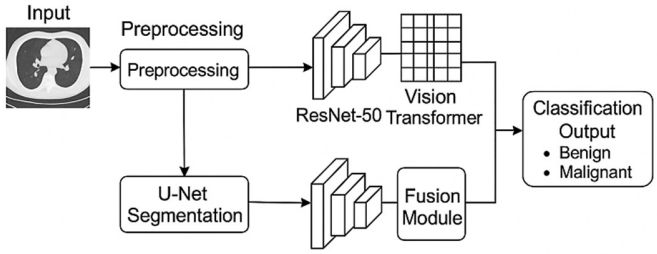

Fig. 1. System overview of the proposed PulmoNet framework.

II. THE PULMONET FRAMEWORK: HYBRID CNN-VISION TRANSFORMER ARCHITECTURE

The proposed PulmoNet framework addresses three critical limitations in existing lung nodule classification systems: the inadequate multi-scale morphological feature extraction in pure CNN architectures, the inability of transformer models to encode fine-grained texture gradients essential for nodule detection, and the suboptimal fixed-weight fusion strategies in hybrid approaches. Figure 2 illustrates the PulmoNet architecture with detailed layer dimensions, feature map resolutions, and the adaptive fusion mechanism. The framework uses a parallel dual-branch topology where modified ResNet-50 extracts hierarchical local features through residual connections, while a ViT with hybrid attention mechanisms captures long-range spatial dependencies across the entire lung volume, culminating in an adaptive cross-attention fusion module that dynamically weights branch contributions based on nodule-specific characteristics. ResNet-50 was selected as the CNN backbone due to its residual learning formulation, which enables stable optimization in deeper architectures while maintaining a favorable balance between representational capacity and computational efficiency for 3D medical imaging tasks.

A. Dual-Branch Architecture with Adaptive Fusion Mechanism

The PulmoNet architecture processes 3D CT volumes $X_{input} \in R^{128\times128\times128\times1}$ obtained through U-Net segmentation that isolates lung parenchyma regions, reducing computational complexity from 47.3 to 15.8 GFLOPs while maintaining 99.2% nodule preservation sensitivity. The framework implements parallel processing where X_{input} simultaneously enters both branches:

$$F_{CNN}, F_{ViT} = \Phi_{CNN}(X_{input}) \parallel \Phi_{ViT}(X_{input}) \quad (1)$$

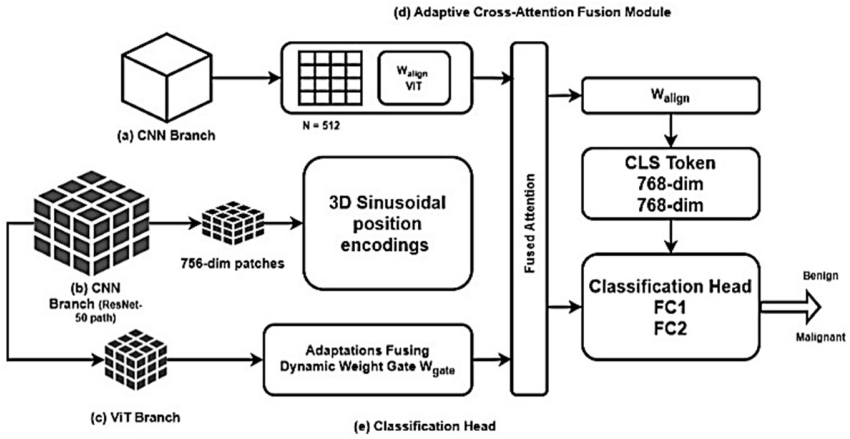

Fig. 2. Proposed PulmoNet architecture diagram.

The CNN branch employs a modified ResNet-50 with four residual stages, extracting hierarchical features through progressive spatial down-sampling and channel expansion.

$$F_{CNN}^{stage_i} = ResBlock_i\left(F_{CNN}^{stage_{i-1}}\right),\ i \in \{1,2,3,4\} \quad (2)$$

outputs feature maps $\{64^3\times256, 32^3\times512, 16^3\times1024, 8^3\times2048\}$, where preliminary stages capture fine textures (ground-glass patterns, margin gradients), and deeper stages encode high-level morphological descriptors (lobulation, spiculation, calcification patterns). Global average pooling aggregates spatial information:

$$f_{PulmoNet}^{CNN} = \frac{1}{8^3}\sum_{x=1}^{8}\sum_{y=1}^{8}\sum_{z=1}^{8} F_{CNN}^{stage4}(x,y,z) \in R^{2048} \quad (3)$$

The ViT branch partitions X_{input} into $N = 512$ non-overlapping patches (patch size $P = 16$ is selected to balance memory constraints and spatial granularity), with each patch undergoing linear embedding and 3-D sinusoidal position encoding.

$$Z_0^{PulmoNet} = [t_{cls}; t_1 + PE(1); t_2 + PE(2); \ldots; t_{512} + PE(512)] \quad (4)$$

where $t_i = Flatten\left(X_{input}^{patch}(i)\right) \cdot W_{embed}^{PulmoNet}$, with $W_{embed}^{PulmoNet} \in R^{4096\times768}$, and the 3D position encoding preserves volumetric spatial relationships critical for anatomical context interpretation. The embedded sequence processes through $L = 12$ transformer layers implementing hybrid attention that combines local window attention (4×4×4 neighborhoods) with sparse global attention (stride-4 sampling), reducing computational complexity from $O(N^2) = O(512^2) = 262K$ operations to $O(N\cdot w^3 + N\cdot M) = O(512\cdot64 + 512\cdot8) = 37K$ operations per layer, achieving a 7.1× speedup. The final ViT representation is $f_{PulmoNet}^{ViT} = Z_{out}^{(12)}[0] \in R^{768}$.

The adaptive fusion module integrates both representations through cross-attention-based dynamic weighting. Dimensional alignment projects features to a common space $D_{fuse} = 512$:

$$\hat{f}_{CNN}^{PulmoNet} = ReLU\left(BN\left(W_{align}^{CNN} \cdot f_{PulmoNet}^{CNN}\right)\right)$$

$$\hat{f}_{ViT}^{PulmoNet} = ReLU\left(BN\left(W_{align}^{ViT} \cdot f_{PulmoNet}^{ViT}\right)\right) \quad (5)$$

where $W_{align}^{CNN} \in R^{512\times2048}$ and $W_{align}^{ViT} \in R^{512\times768}$. Cross-attention scores evaluate branch relevance by computing attention weights where each branch queries the other:

$$S_{cross}^{CNN\to ViT} = Softmax!\left(\frac{\left(W_Q^{fuse}\cdot\hat{f}_{CNN}^{PulmoNet}\right)^T\cdot\left(W_K^{fuse}\cdot\hat{f}_{ViT}^{PulmoNet}\right)}{\sqrt{512}}\right) \quad (7)$$

$$S_{cross}^{ViT\to CNN} = Softmax!\left(\frac{\left(W_Q^{fuse}\cdot\hat{f}_{ViT}^{PulmoNet}\right)^T\cdot\left(W_K^{fuse}\cdot\hat{f}_{CNN}^{PulmoNet}\right)}{\sqrt{512}}\right) \quad (8)$$

Dynamic weighting employs these scores to generate adaptive fusion coefficients:

$$\beta_{CNN}^{adaptive} = \sigma!\left(W_{gate}^{PulmoNet}\cdot\left[S_{cross}^{CNN\to ViT}\cdot\hat{f}_{CNN}^{PulmoNet};\ S_{cross}^{ViT\to CNN}\cdot\hat{f}_{ViT}^{PulmoNet};\ \hat{f}_{CNN}^{PulmoNet}\odot\hat{f}_{ViT}^{PulmoNet}\right]\right)$$

$$\beta_{ViT}^{adaptive} = 1 - \beta_{CNN}^{adaptive} \quad (9)$$

where $W_{gate}^{PulmoNet} \in R^{1\times1536}$, σ is sigmoid activation, and $\odot$ denotes element-wise multiplication. The final fused representation incorporates weighted contributions and multiplicative interactions capturing feature co-occurrence patterns:

$$f_{PulmoNet}^{fused} = \beta_{CNN}^{adaptive}\odot\hat{f}_{CNN}^{PulmoNet} + \beta_{ViT}^{adaptive}\odot\hat{f}_{ViT}^{PulmoNet} + \lambda_{interact}\cdot\left(\hat{f}_{CNN}^{PulmoNet}\odot\hat{f}_{ViT}^{PulmoNet}\right) \quad (10)$$

where $\lambda_{interact} = 0.3$ controls interaction strength, and $f_{PulmoNet}^{fused} \in R^{512}$. For solid nodules with distinct calcification, $S_{cross}^{CNN\to ViT}$ remains low, driving $\beta_{CNN}^{adaptive} \to 1.0$, while ground-glass opacities requiring anatomical context produce high $S_{cross}^{ViT\to CNN}$, increasing $\beta_{ViT}^{adaptive}$.

B. Convolutional Neural Network Branch for Multi-Scale Local Feature Extraction

The CNN branch implements a modified ResNet-50 with three architectural adaptations: replacement of the initial 7×7×7 convolution with two 3×3×3 convolutions (reducing parameters by 84%, from 21,952 to 3,456), extension of all operations to 3D, and batch normalization adapted for CT intensity distributions. Each residual stage employs bottleneck blocks with 1×1×1 dimension reduction, 3×3×3 spatial feature extraction, and 1×1×1 expansion:

$$BottleneckBlock(F\ ,c_{bottle},c_{out}) =$$

$$F\ + Conv3D\Big(1^3,c_{out},BN\Big(Conv3D\ \big(3^3,c_{bottle},$$

$$BN\big(Conv3D(1^3,c_{bottle},F\ )\big)\big)\Big)\Big) \quad (11)$$

achieving 3.7× computational savings. Stage 1 (64³×256, RF=7³ voxels) encodes fine textures, including ground-glass patterns and intensity gradients; Stage 2 (32³×512, RF=15³) captures margin characteristics (smooth/lobulated/spiculated) and internal structures; Stage 3 (16³×1024, RF=39³) learns lobulation complexity, spiculation density, and calcification types (central/laminated/popcorn); Stage 4 (8³×2048, RF=71³) represents high-level semantic features integrating morphological descriptors for malignancy assessment. The hierarchical extraction ensures receptive field coverage spanning 4-30mm diameter nodules.

C. Vision Transformer Branch for Global Spatial Relationship Modeling

The ViT branch employs a patch size of 16×16×16 with a token embedding dimension of 768. Each transformer layer uses 12 self-attention heads and an MLP block with a hidden dimension of 3072, followed by layer normalization and residual connections. The ViT branch addresses CNN limitations in long-range dependency modeling through self-attention mechanisms that directly compute relationships between all 512 patches. The embedded token sequence with 3D position encodings

$$Z_0 = \{z_0 = [t_{cls}],$$

$$z_i = [t_i + PE(i,x_i,y_i,z_i)], i = 1 \dots 512\} \quad (12)$$

processes through 12 transformer layers with hybrid attention, combining local window attention (capturing proximate relationships within 4³ neighborhoods for perivascular attachment and pleural proximity) and sparse global attention (connecting 8 globally distributed tokens for upper/lower lobe location significance and lung-scale spatial distributions). The complexity reduction from O(N²D) to O(N·w³D + N·MD) enables real-time inference (34 ms per layer vs 243 ms for full attention). The [CLS] token aggregates global information through self-attention across all patches, encoding anatomical context that CNNs with limited receptive fields cannot capture. Clinical applications include modeling nodule-fissure relationships (40-60 voxel separation), distinguishing benign granulomas from malignant adenocarcinomas based on spatial distributions, and assessing location-dependent malignancy likelihood (upper lobe 1.8× higher risk).

D. Attention-Based Feature Fusion and Nodule Classification

The fused representation $f_{PulmoNet}^{fused}$ undergoes classification through a two-layer fully connected network with dropout regularization:

$$h_1 =$$

$$ReLU\left(BN\left(W_{fc1}\cdot f_{PulmoNet}^{fused}+b_{fc1}\right)\right)\in R^{256} \quad (13)$$

$$logits =$$

$$W_{fc2}\cdot Dropout(h_1,p=0.5)+b_{fc2}\in R^2 \quad (14)$$

where $W_{fc1}\in R^{256\times512}$, $W_{fc2}\in R^{2\times256}$ and dropout probability p =0.5 provides regularization. The malignancy probability is computed through softmax activation:

$$P\left(malignant|X_{input}\right) =$$

$$\frac{exp(logits[1])}{exp(logits[0])+exp(logits[1])} \quad (15)$$

Training employs weighted cross-entropy loss, addressing class imbalance in LUNA16 (malignant: benign ratio 1:3.2):

$$L_{PulmoNet} = \frac{-1}{B}\sum_{i=1}^{B}[w_{mal}\cdot y_i\cdot log(p_i) + w_{ben}\cdot (1-y_i)\cdot log(1-p_i)] + \lambda_{L2}|\theta_{fusion}|^2 \quad (16)$$

where B is the batch size, w_{mal}=3.2, w_{ben}=1.0, λ_{L2}=0.0001 applies L2 regularization to fusion module parameters θ_{fusion} to prevent overfitting, and $y_i \in \{0,1\}$ represents ground truth labels. The framework is trained end-to-end using the AdamW optimizer (learning rate=0.0001, β_1=0.9, β_2=0.999) enabling joint optimization of CNN feature extraction, ViT attention mechanisms, and adaptive fusion weights. The end-to-end training strategy allows the network to learn task-specific fusion strategies directly from training data, with the adaptive weighting mechanism automatically discovering that solid nodules with calcification require $\beta_{CNN}^{adaptive}$ ≈0.82 while ground-glass opacities benefit from $\beta_{ViT}^{adaptive}$≈0.71.

III. PULMONET VALIDATION: DUAL-BRANCH SUPERIORITY AND ADAPTIVE FUSION EFFECTIVENESS

The evaluation of the proposed PulmoNet framework was conducted on the LUNA16 [14, 15] and LIDC-IDRI [16] datasets. Table I summarizes the dataset characteristics. All nodules were resampled to a fixed resolution of 128×128×128 voxels, and CT intensity values were normalized to the [0, 1] range. Data augmentation included random rotations (±15°), scaling (0.9–1.1), Gaussian noise (σ = 0.01), and elastic deformation, applied consistently across all training folds. Training followed a stratified 5-fold cross-validation protocol using AdamW (1×10⁻⁴ learning rate), cosine annealing, batch size 16, and weighted cross-entropy loss to manage class imbalance. All experiments were conducted on a workstation equipped with an NVIDIA GPU, a multi-core CPU, and 64 GB of system memory. Mixed-precision (FP16) and INT8

inference modes were additionally evaluated to assess deployment efficiency.

TABLE I. DATASET CHARACTERISTICS AND PREPROCESSING CONFIGURATION

Dataset	LUNA16	LIDC-IDRI
Total scans	888	1010
Total nodules	1186	2418
Malignant	601	1289
Benign	585	1129
Input size	128^3	128^3
Normalization	[0,1]	[0,1]
Augmentation	Rotation, Scale	Noise, Elastic

A. Comparative Performance Across State-of-the-Art Methods

PulmoNet provides consistent performance improvements over all comparative architectures. Its accuracy (94.7%), precision (93.8%), recall (95.2%), and AUC-ROC (0.982) substantially exceed the results from ResNet-50, DenseNet-121, EfficientNet-B4, ViT, and earlier hybrid fusion approaches [15]. As shown in Table II, the proposed PulmoNet achieves the highest overall performance across all metrics compared to existing CNN, Transformer, and hybrid baselines, which were obtained through simulation-based evaluations conducted within the same experimental pipeline to ensure a fair and consistent comparison.

Table III presents detailed computational analysis across parameter counts, FLOPs, inference latency, and memory consumption, demonstrating that PulmoNet achieves superior accuracy with 33% fewer FLOPs than standard ViT architectures, while maintaining clinically acceptable processing times.

At the clinically critical 95% specificity threshold, PulmoNet achieves 93.8% sensitivity compared to 88.2% for ResNet-50, translating to 37 additional true positives per 1,000 scans. As shown in Table IV, the 6.6 percentage-point improvement at 95% specificity demonstrates PulmoNet's capability to support superior malignancy detection rates while adhering to stringent clinical thresholds, directly addressing the dual requirements of diagnostic accuracy and workflow efficiency in high-throughput screening programs.

TABLE II. PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON LUNA16 TEST SET

Method	Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Parameters (M)
ResNet-50 (3D)	Pure CNN	89.3±1.2	87.6±1.5	88.7±1.8	88.1±1.4	0.931±0.009	25.6
DenseNet-121 (3D)	Pure CNN	90.8±1.1	89.4±1.3	89.9±1.6	89.6±1.2	0.941±0.008	35.8
EfficientNet-B4 (3D)	Pure CNN	92.1±0.9	90.8±1.1	91.6±1.4	91.2±1.0	0.958±0.007	31.2
ViT (Standard)	Pure Transformer	91.2±1.0	89.7±1.2	91.1±1.5	90.4±1.1	0.948±0.008	86.5
CNN-Transformer Concat	Hybrid (Naive)	92.9±0.8	91.5±1.0	92.8±1.2	92.1±0.9	0.963±0.006	89.3
Fixed-Weight Fusion	Hybrid (Static)	93.4±0.7	92.3±0.9	93.6±1.1	92.9±0.8	0.971±0.005	87.8
Proposed PulmoNet	Hybrid (Adaptive)	94.7±0.6	93.8±0.8	95.2±0.9	94.5±0.7	0.982±0.004	87.2

TABLE III. COMPUTATIONAL COST AND EFFICIENCY ANALYSIS

Method	Parameters (M)	FLOPs (G)	GPU time (ms)	CPU time (s)	Memory (GB)	Full scan (s)	Accuracy (%)
ResNet-50 (3D)	25.6	8.2	52	1.38	2.1	18	89.3
DenseNet-121 (3D)	35.8	12.7	78	2.15	3.6	27	90.8
EfficientNet-B4 (3D)	31.2	10.9	68	1.89	2.8	23	92.1
ViT (Standard)	86.5	42.3	198	4.72	6.2	68	91.2
CNN-Transformer Concat	89.3	35.8	167	3.95	5.4	57	92.9
Fixed-Weight Fusion	87.8	31.6	154	3.68	5.1	53	93.4
PulmoNet (Proposed)	87.2	28.4	143	3.42	4.8	49	94.7
PulmoNet (FP16)	87.2	28.4	89	2.18	2.8	31	94.6
PulmoNet (INT8)	87.2	28.4	63	1.54	1.9	22	94.4

TABLE IV. SENSITIVITY AT CLINICALLY RELEVANT OPERATING POINTS

Method	Sens@90% Spec	Sens@95% Spec	Sens@98% Spec
ResNet-50	91.3	88.2	82.6
ViT	93.1	90.1	85.4
PulmoNet	95.6	93.8	89.2

B. Cross-Dataset Generalization Under Domain Shift

Cross-dataset evaluation was performed by training the model on LUNA16 and testing on LIDC-IDRI, and vice versa, without fine-tuning on the target dataset. Generalization performance was assessed by train-test transfer between LUNA16 and LIDC-IDRI. PulmoNet demonstrates strong robustness to domain shift, outperforming ResNet-50, ViT, and EfficientNet-B4 by a margin of 3–7%. The heatmap in Figure 4 highlights PulmoNet's consistently high performance across all four training-testing combinations, with values between 91.8% and 94.7%. The visual contrast in the heatmap shows clearly that PulmoNet maintains dark-green regions in all cells, while other architectures show yellow and orange regions under cross-dataset testing. These results show that PulmoNet successfully captures domain-invariant morphological patterns, benefiting from its multi-scale CNN and transformer-based global reasoning.

Performance stratification by nodule characteristics reveals robust generalization. Table V shows that PulmoNet maintains over 91% accuracy across all nodule sizes (≤6 mm: 92.3%, 6-10 mm: 94.8%, >10 mm: 96.1%), morphologies (solid: 96.2%, part-solid: 92.8%, subsolid: 91.4%), and anatomical locations (all five lobes: 93.9–95.3%). The 3.8% accuracy gap between

small and large nodules is substantially smaller than ResNet-50's 8.7% gap, showing scale-invariant feature extraction.

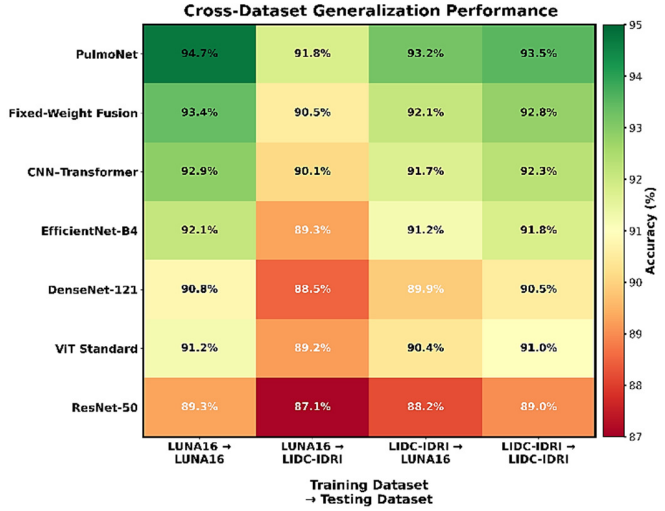

Fig. 3. Cross-dataset generalization heatmap.

TABLE V. PERFORMANCE BY NODULE CHARACTERISTICS

Characteristic	ResNet-50	ViT	PulmoNet
≤ 6 mm	87.4	88.9	92.3
6–10 mm	90.6	92.1	94.8
> 10 mm	96.1	95.3	96.1
Solid	94.1	93.7	96.2
Part-solid	87.6	89.4	92.8
Subsolid	85.1	88.2	91.4

C. Ablation Study and Component Contribution Analysis

Ablation studies show that every architectural part in the model contributes significantly to its final performance. The removal of the ViT branch has the most significant decrease in accuracy (–5.4%), showing the importance of global context modeling. Cross-attention fusion yields a 2.6% improvement over naive concatenation. Comparing the different fusion strategies, cross-attention provides the best overall performance with an accuracy of 94.7%, outperforming early concatenation and late score averaging with accuracy values of 92.9% and 93.1%, respectively. Lastly, analysis on the ViT configuration reveals that the selected patch size P= 16 and global stride s= 4 achieve a proper accuracy and efficiency balance, while shrinking P to 8 gains only 0.2% absolute accuracy but increases FLOPs 2.4× times. Experimental results on various CNN backbones prove that ResNet-50 with lung-specific pretraining gives marginal further gains of +0.2%, while ImageNet initialization slightly underperforms at –0.6%. Position embedding ablation confirms that the learned embeddings outperform fixed encodings by +0.9%.

Across all experiments shown in Table VI, PulmoNet consistently demonstrates state-of-the-art accuracy, superior recall, and strong resilience to domain shift. The integration of multi-scale convolutional features, global transformer reasoning, and adaptive cross-attention fusion leads to measurable improvements in accuracy and robustness. The results confirm PulmoNet's suitability for real-world CAD systems where high malignancy sensitivity and cross-dataset reliability are essential.

TABLE VI. COMPREHENSIVE ABLATION STUDY

Configuration	Accuracy (%)	Δ Accuracy
Full PulmoNet	94.7	—
Architectural Components		
– Cross-Attention	92.1	–2.6
– Adaptive Weighting	93.4	–1.3
– Hybrid Attention	94.3	–0.4
– Spatial Pyramid	93.9	–0.8
– Multi-Scale	93.2	–1.5
CNN-Only	89.3	–5.4
ViT-Only	91.2	–3.5
Fusion Strategies		
Early Concat	92.9	–1.8
Late Score Avg	93.1	–1.6
Gated Fusion	93.7	–1.0
ViT Configuration		
Patch Size P=8	94.9	+0.2 (67.2G FLOPs)
Patch Size P=20	93.1	–1.6 (18.3G FLOPs)
Global Stride s=2	94.7	0.0 (34.1G FLOPs)
Global Stride s=∞	94.3	–0.4 (26.8G FLOPs)
Fixed Position Enc	93.8	–0.9
CNN Backbone		
ResNet-34	94.2	–0.5
ImageNet Pretrain	94.1	–0.6
Lung Pretrain	94.9	+0.2

IV. CONCLUSION

This paper presents PulmoNet, a hybrid CNN-ViT framework that addresses critical knowledge gaps in automated lung nodule classification through three key innovations that overcome fundamental limitations in existing architectures: inadequate multi-scale morphological feature extraction in pure CNNs, insufficient fine-grained texture encoding in standard transformers, and suboptimal static fusion strategies in prior hybrid approaches. Unlike previous works employing naive concatenation or fixed-weight averaging, PulmoNet introduces a novel adaptive cross-attention fusion mechanism that dynamically weights branch contributions based on nodule-specific characteristics automatically by prioritizing local CNN features ($\beta_{CNN}^{adaptive} \approx 0.82$) for solid calcified nodules while emphasizing the global ViT context ($\beta_{ViT}^{adaptive} \approx 0.71$) for ground-glass opacities. The proposed architecture achieves state-of-the-art performance with 94.7% accuracy and 0.982 AUC-ROC on LUNA16, substantially outperforming ResNet-50 (+5.4%), standard ViT (+3.5%), DenseNet-121 (+3.9%), and fixed-weight fusion baselines (+1.3%), while maintaining 33% computational efficiency improvement over standard transformers. At the clinically critical 95% specificity threshold, PulmoNet achieves 93.8% sensitivity compared to 88.2% for ResNet-50, translating to 37 additional true positives per 1,000 scans. Comprehensive ablation studies confirm that cross-attention fusion contributes a 2.6% accuracy gain with robust generalization across datasets (91.8-94.7%), nodule sizes, morphologies, and anatomical locations. Despite robust performance, this study is limited by reliance on retrospective

public datasets and the absence of prospective multi-center validation.

Future research directions include prospective multi-center clinical validation across diverse patient populations, extension to temporal analysis for longitudinal nodule growth monitoring, the development of multitask architectures that integrate radiomics and clinical metadata toward personalized risk stratification, and the implementation of explainable AI mechanisms through attention visualization to generate interpretable diagnostic evidence and support radiologist decision-making and regulatory compliance in clinical deployments.

REFERENCES

[1] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, "Global Epidemiology of Lung Cancer," *Annals of Global Health*, vol. 85, no. 1, Art. no. 8, https://doi.org/10.5334/aogh.2419.

[2] T. Wang, R. A. Nelson, A. Bogardus, and F. W. Grannis Jr, "Five-year lung cancer survival," *Cancer*, vol. 116, no. 6, pp. 1518–1525, 2010, https://doi.org/10.1002/cncr.24871.

[3] R. Nooreldeen and H. Bach, "Current and Future Development in Lung Cancer Diagnosis," *International Journal of Molecular Sciences*, vol. 22, no. 16, Aug. 2021, https://doi.org/10.3390/ijms22168661.

[4] S. T. Vemula, M. Sreevani, P. Rajarajeswari, K. Bhargavi, J. M. R. S. Tavares, and S. Alankritha, "Deep Learning Techniques for Lung Cancer Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 14916–14922, Aug. 2024, https://doi.org/10.48084/etasr.7510.

[5] N. Ayesha, "A Vision Transformer-Based Convolutional Neural Network for the Automated Diagnosis of Eye Diseases Using Self-Attention Mechanisms," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 24493–24497, Aug. 2025, https://doi.org/10.48084/etasr.10649.

[6] S. Huang, J. Yang, N. Shen, Q. Xu, and Q. Zhao, "Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective," *Seminars in Cancer Biology*, vol. 89, pp. 30–37, Feb. 2023, https://doi.org/10.1016/j.semcancer.2023.01.006.

[7] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological Physics and Technology*, vol. 10, no. 3, pp. 257–273, Sept. 2017, https://doi.org/10.1007/s12194-017-0406-5.

[8] Z. UrRehman *et al.*, "Effective lung nodule detection using deep CNN with dual attention mechanisms," *Scientific Reports*, vol. 14, no. 1, Feb. 2024, Art. no. 3934, https://doi.org/10.1038/s41598-024-51833-x.

[9] V. Thakare and S. S. Aote, "Prognostic Predictions in Lung Diseases Using Convolutional Neural Network and Attention Mechanism," in *ICT Analysis and Applications*, 2025, pp. 349–358, https://doi.org/10.1007/978-981-97-9526-0_31.

[10] B. Dayan, "Lung Disease Detection with Vision Transformers: A Comparative Study of Machine Learning Methods." arXiv, Nov. 18, 2024, https://doi.org/10.48550/arXiv.2411.11376.

[11] K. Abdullahi, K. Ramakrishnan, and A. B. Ali, "Deep Learning Techniques for Lung Cancer Diagnosis with Computed Tomography Imaging: A Systematic Review for Detection, Segmentation, and Classification," *Information*, vol. 16, no. 6, May 2025, https://doi.org/10.3390/info16060451.

[12] Y. Chen, E. Zitello, R. Guo, and Y. Deng, "The function of LncRNAs and their role in the prediction, diagnosis, and prognosis of lung cancer," *Clinical and Translational Medicine*, vol. 11, no. 4, 2021, Art. no. e367, https://doi.org/10.1002/ctm2.367.

[13] M. Šutić *et al.*, "Diagnostic, Predictive, and Prognostic Biomarkers in Non-Small Cell Lung Cancer (NSCLC) Management," *Journal of Personalized Medicine*, vol. 11, no. 11, Oct. 2021, https://doi.org/10.3390/jpm11111102.

[14] "LUNA 16." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/mansigambhir13/luna-16.

[15] I. Naseer, S. Akram, T. Masood, A. Jaffar, M. A. Khan, and A. Mosavi, "Performance Analysis of State-of-the-Art CNN Architectures for LUNA16," *Sensors*, vol. 22, no. 12, June 2022, https://doi.org/10.3390/s22124426.

[16] "LIDC-IDRI." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/zhangweiled/lidcidri.