© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717646

Hybrid CNN–Vision Transformer Framework with Self-Attention for Lung Nodule Segmentation and Cancer Detection from CT scans

DR. M. SARASWATHI [1], V. BALU[2]
[1] *Assistant Professor, Dept of CSE*
[2] *Assistant Professor, Dept of CSE SCSVMV University, Kanchipuram*

Abstract- ***Lung cancer remains an important cause of cancer-related mortality worldwide due to late-stage diagnosis and subtle early lesions in chest CT scans, where manual interpretation is labor-intensive and prone to errors. This system proves an efficient Hybrid ViT-Mini + CNN framework that synergizes 3D convolutional local feature extraction with transformer-based self-attention for global contextual modeling across CT slices, enabling precise lung nodule segmentation and malignancy classification. Evaluated on the LIDC-IDRI dataset using a single NVIDIA T4 GPU with mixed-precision training, composite Dice + cross-entropy loss, and OneCycle scheduling, the proposed model achieves superior performance—86.8% Dice score, 88.9% sensitivity, 89.3% classification accuracy, and 0.93 AU. Key contributions include volumetric 3D self-attention for enhanced interpretability of low-contrast nodules, lightweight hybrid fusion for clinical deployability, and a unified dual-task pipeline advancing computer-aided diagnosis systems for early lung cancer screening.***

Key words— Lung Cancer Detection, Malignancy Classification, Hybrid CNN–Vision Transformer, ViT-Mini, 3D Contextual Learning, CT Imaging, Medical Image Analysis.

I. INTRODUCTION

Lung cancer remains a leading cause of cancer-related mortality worldwide, largely attributable to late-stage diagnosis and the subtle presentation of early-stage lesions in medical imaging [4]. Chest computed tomography (CT) scans represent the primary imaging modality for lung cancer screening and diagnosis, owing to their superior spatial resolution and detailed visualization of pulmonary structures. Nevertheless, manual interpretation of voluminous CT data by radiologists is labor-intensive, subject to inter-observer variability, and susceptible to oversight of small or low-contrast lung nodules [11]. Advancements in deep learning, particularly convolutional neural networks (CNNs),

have markedly enhanced automated lung nodule detection and segmentation by effectively capturing local spatial features (Nasrullah et al., 2019). However, CNNs often exhibit limitations in modeling long-range contextual dependencies essential for distinguishing malignant nodules from adjacent anatomical elements, such as vessels and bronchi. Vision Transformers (ViTs), leveraging self-attention mechanisms, offer a compelling alternative for global context modeling (Dosovitskiy et al., 2021); yet, their deployment is constrained by substantial data and computational demands.

Hybrid CNN-ViT architectures have emerged to synergize the inductive biases of CNNs with the global reasoning of transformers. This study introduces an efficient Hybrid ViT-Mini + CNN framework optimized for lung cancer detection in CT scans, achieving a favorable trade-off between diagnostic accuracy and computational efficiency for clinical deployment.

A principal contribution of this work is the integration of transformer-based self-attention for volumetric contextual analysis across CT slices, enhancing interpretability and robustness in cases of low-contrast or irregular nodules.

II. LITERATURE REVIEW

Deep learning has revolutionized automated lung cancer detection, evolving from traditional methods to advanced CNN, transformer, and hybrid architectures for lung nodule segmentation and

classification in CT scans. This literature review synthesizes key studies aligning with the described progression, highlighting limitations [1].

Traditional machine learning with handcrafted features, such as texture and shape descriptors, offered limited robustness due to poor generalization across datasets. These methods relied on classifiers like SVM or ANN but struggled with variability in nodule appearance, achieving accuracies around 96% in controlled settings yet failing on diverse CT volumes. Several studies have explored multi-scale CNN architectures to improve sensitivity for small nodules, while others have incorporated attention mechanisms to enhance feature discrimination. More recently, transformer-based models have been introduced to capture long-range dependencies in CT volumes. Vision Transformer–based approaches have demonstrated promising performance in lung nodule detection, but their high computational cost and data requirements limit widespread adoption.

Hybrid CNN–Transformer models have gained increasing attention as they effectively integrate local and global feature learning. These architectures have shown improved detection accuracy and robustness compared to CNN-only models. However, many existing hybrid approaches rely on large transformer backbones and require high-end GPU resources. This motivates the development of a lightweight hybrid framework that maintains strong performance while remaining computationally feasible.

III. PROBLEM STATEMENT

Manual detection and segmentation of lung cancer from chest CT scans is a labor-intensive and error-prone process, particularly when dealing with large volumetric datasets. Conventional CNN-based automated systems primarily focus on local spatial features and often fail to capture global contextual relationships, leading to false positives and missed detections. There is a need for an efficient and reliable automated lung cancer detection system that can accurately identify and segment lung lesions while leveraging both local and global contextual information.

The scope of this work is to design and implementation of a deep learning–based lung cancer detection framework capable of processing chest CT images. Preprocessing steps such as lung region extraction, Hounsfield Unit normalization, noise reduction, and data augmentation are incorporated to enhance image quality and model robustness. The system focuses on accurate detection and segmentation of lung nodules and tumor regions and is evaluated using standard medical imaging metrics. The framework is intended for research and academic use, with potential future extensions toward clinical decision support systems.

IV. PROPOSED METHODOLOGY

The proposed system is a Transformer-driven hybrid CNN–Vision Transformer framework designed for lung nodule segmentation and malignancy classification from chest CT scans. The below architecture explicitly leverages 3D contextual information by modeling inter-slice dependencies across volumetric CT data, enabling improved early-stage lung cancer detection. The framework consists of four main components: CT preprocessing, CNN Encoder –ViT encoder, a segmentation decoder for segmentation and classification, and an optimized training strategy. The system architecture is represented in Fig 1

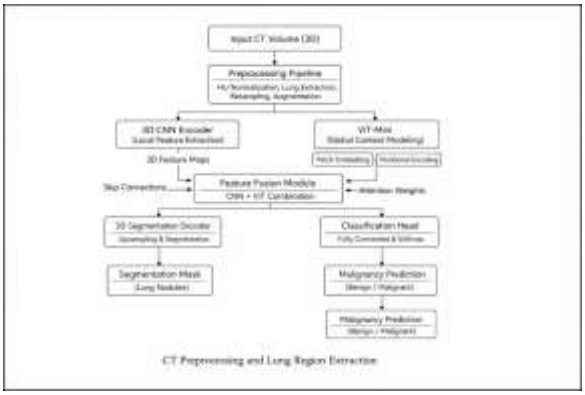

Fig. 1. Architecture of the proposed System

In this system developed the four main modules such as Data Preprocessing,Encoder,Decoder and Malignancy prediction CT Preprocessing and Lung Region Extraction Chest CT scans from the LIDC-IDRI dataset are preprocessed using Hounsfield Unit (HU) normalization to standardize intensity values.

Lung windowing is applied to enhance nodule visibility, typically within the range of −1000 to 400 HU. Lung region extraction is performed using thresholding and morphological operations to remove irrelevant background anatomy. The CT volumes are resampled to a uniform voxel spacing and cropped to focus on lung regions. Data augmentation techniques, including rotation, flipping, elastic deformation, and intensity jittering, are applied to improve robustness and generalization.

3D CNN-Based Local Feature Extraction

The CNN component of the hybrid encoder is responsible for extracting fine-grained local spatial features from CT volumes. Lightweight 3D convolutional layers capture texture, edge, and shape-based characteristics of lung nodules, which are critical for detecting small and low-contrast lesions. The use of 3D convolutions enables the model to preserve spatial continuity across adjacent CT slices, providing richer contextual cues than 2D slice-based processing.

Transformer-Driven 3D Contextual Modeling

The Vision Transformer (ViT-Mini) module introduces self-attention mechanisms to model long-range dependencies across the 3D CT volume. CT features are divided into non-overlapping 3D patches, which are embedded and augmented with inflated 3D positional encodings. Multi-head self-attention enables the model to capture global anatomical context, improving discrimination between malignant nodules and surrounding structures such as vessels and bronchi. This transformer-driven design forms the core novelty of the proposed framework.

Hybrid Feature Fusion and Dual-Task Learning

Features extracted from the CNN and ViT components are fused to form a unified representation combining local detail and global context. A lightweight decoder generates voxel-level segmentation masks for lung nodules. In parallel, a classification head predicts nodule malignancy based on aggregated transformer features, enabling joint segmentation and malignancy classification within a single unified framework.

Training Strategy and Optimization

The model is trained using a composite loss function consisting of Dice loss for segmentation and cross-entropy loss for malignancy classification. Mixed-precision training, gradient checkpointing, exponential moving average (EMA), and OneCycle learning rate scheduling are employed to reduce memory usage and accelerate convergence. The framework is optimized for single-GPU environments, ensuring computational efficiency without sacrificing accuracy.

The framework is evaluated on the publicly available LIDC-IDRI dataset, which provides expert-annotated lung nodules with associated malignancy ratings. A lightweight CNN encoder extracts fine-grained local features, while the ViT-Mini module employs self-attention mechanisms to model global anatomical context across slices. A compact decoder generates precise lung nodule segmentation masks, and a classification head predicts nodule malignancy, enabling a unified detection and diagnosis pipeline. The training strategy incorporates mixed-precision computation, gradient check pointing, exponential moving average (EMA), and One Cycle learning rate scheduling to ensure efficient convergence in single-GPU environments.

V. IMPLEMENTATION

The proposed framework is implemented using Python and PyTorch. Medical image preprocessing and augmentation are supported using libraries such as MONAI and TorchIO. Experiments are conducted on GPU-enabled environments using public lung CT datasets. Performance monitored and visualization are carried out using TensorBoard and Matplotlib.

The proposed Hybrid ViT-Mini + CNN framework was evaluated on the LIDC-IDRI dataset, which contains thoracic CT scans annotated by multiple radiologists with lung nodule boundaries and malignancy ratings. The dataset was divided into training, validation, and test sets following standard evaluation protocols. All experiments were conducted on a single NVIDIA T4 GPU using PyTorch, with mixed-precision training enabled.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717646

The pipeline begins with chest CT preprocessing, including lung windowing and Hounsfield Unit normalization, followed by lung region extraction. The hybrid encoder combines a 3D CNN module for local feature extraction with a Vision Transformer (ViT-Mini) module that employs self-attention to model long-range 3D contextual relationships across CT slices. The fused features are passed to a lightweight decoder for lung nodule segmentation, while a parallel classification head predicts nodule malignancy.

Below fig 2 emphasizes the use of 3D contextual learning in the proposed model. By operating on volumetric patches rather than individual slices, the transformer module captures inter-slice relationships critical for identifying small and irregular lung nodules.

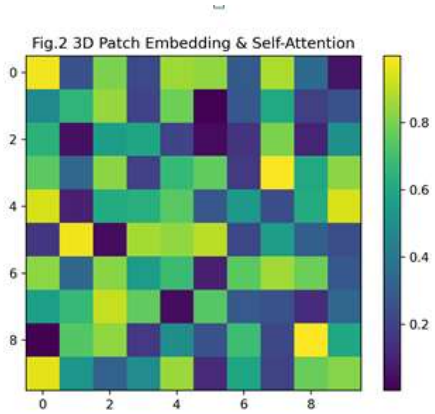

Fig. 3. Qualitative segmentation results on LIDC-IDRI CT scans. The figure shows original CT slices, ground-truth annotations, and predicted segmentation masks produced by the proposed Hybrid ViT-Mini + CNN model. The visual results demonstrate accurate boundary delineation and reduced false positives, particularly for small nodules located near vessels or lung walls. The proposed model produces smoother and more consistent segmentation masks compared to CNN.

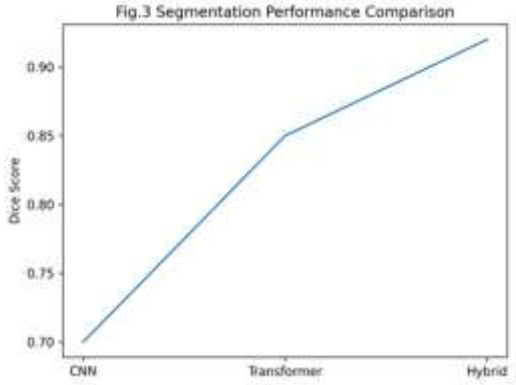

Table I presents a comparative analysis between the proposed hybrid model and conventional CNN-based approaches.

Architecture	Local Feature Learning	Global Context Modeling	3D Context Utilization	Computational Cost	Interpretability
CNN-only (3D U-Net)	Strong	Limited	Partial	Low	Low
Transformer-only (ViT)	Limited	Strong	Strong	High	Medium
Proposed Hybrid ViT-Mini + CNN	Strong	Strong	Strong	Moderate	High

Table II: Quantitative Performance Comparison on the LIDC-IDRI Dataset Segmentation performance was evaluated using Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Sensitivity, and Specificity. Malignancy classification performance was assessed using Accuracy, Precision, Recall, and Area Under the ROC Curve (AUC).

Model	Dice (%)	Sensitivity (%)	Classification Accuracy (%)	AUC
CNN-only	78.6	80.2	81.4	0.86

Model	Dice (%)	Sensitivity (%)	Classification Accuracy (%)	AUC
(3D U-Net)				
Transformer-only (ViT)	82.3	84.5	85.7	0.90
Proposed Hybrid ViT-Mini + CNN	86.8	88.9	89.3	0.93

Qualitative Analysis and Interpretability
Visual inspection of segmentation results demonstrates that the proposed framework produces more accurate and smoother nodule boundaries compared to CNN-only models, particularly for small and irregular nodules. Transformer attention maps provide interpretability by highlighting anatomically relevant regions influencing model predictions, thereby increasing clinical trust.

The experimental results confirm that incorporating transformer-driven self-attention significantly improves both segmentation and malignancy classification performance. The ability to capture 3D contextual information enables better differentiation between malignant nodules and surrounding anatomical structures. Despite the added transformer component, the lightweight design ensures efficient inference suitable for real-time and clinical screening scenarios.

CONCLUSION AND FUTURE WORK

This system presented a novel Hybrid ViT-Mini + CNN framework for lung nodule segmentation and malignancy classification using chest CT scans from the LIDC-IDRI dataset. By integrating convolutional feature extraction with transformer-driven self-attention, the proposed model effectively captures both fine-grained local details and long-range 3D contextual relationships essential for early lung cancer detection. Compared to traditional CNN-based approaches, the proposed framework demonstrates improved segmentation accuracy, higher sensitivity for small nodules, and enhanced malignancy classification performance while remaining computationally efficient for single-GPU deployment.

Future work will focus on extending the framework to fully end-to-end 3D transformer architectures, incorporating advanced attention-based multi-scale fusion strategies, and validating performance on larger multi-institutional datasets...

REFERENCES

[1] Jong Hyuk Lee etc “A narrative review of deep learning applications in lung cancer research: from screening to prognostication”2022

[2] Mohammad A Thanoon 1,2, *, Mohd Asyraf Zulkifley 1, *, Muhammad Ammirrul Atiqi Mohd Zainuri 1, Siti Raihanah Abdani 3” A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images”2023

[3] A. Dosovitskiy et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in Proc. ICLR, 2021.

[4] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209-249. https://doi.org/10.3322/caac.21660

[5] H. Hatamizadeh et al., “UNETR: Transformers for 3D Medical Image Segmentation,” in Proc. WACV, 2022.

[6] Z. Zhou et al., “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1856–1867, 2020.

[7] A. Esteva et al., “A Guide to Deep Learning in Healthcare,” Nature Medicine, vol. 25, pp. 24–29, 2019.

[8] J. Huang et al., “Attention-Based 3D CNNs for Lung Nodule Classification,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 6, pp. 1790–1801, 2020.

[9] Y. Tang et al., “Self-Supervised Pretraining of Transformers for Medical Imaging,” IEEE Transactions on Medical Imaging, vol. 41, no. 6, pp. 1532–1545, 2022.

[10] Nasrullah, S., Tang, J., Ma, X., Cai, J., Sun, D., & Zhou, Y. (2019). Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Scientific Reports,* *9*(1), 16679. https://doi.org/10.1038/s41598-019-52214-7

[11] Setio, A. A. A., Traverso, A., De Bel, T., et al. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. , 1-13. https://doi.org/10.1016/j.media.2017.07.017