**Journal of Information Systems Engineering and Management**
2025, 10(15s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Hierarchical Vision Transformer Model-based Lung Cancer Detection with Multiscale Patch Embedding and Cross-Attention Fusion

K Yogeswara Rao[1], K Srinivasa Rao[2]

*1Associate Professor, Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University), India, ykalla@gitam.edu*
*2Associate Professor, Department of Computer Science and Engineering, GITAM School of Technology,GITAM(Deemed to be University),India,ykalla@gitam.edu*

**ARTICLE INFO**

Received: 05 Dec 2024
Revised: 25 Jan 2025
Accepted: 08 Feb 2025

**ABSTRACT**

Lung cancer has become one of the most complex tumors to diagnose early, particularly with CT imaging, due to the intricacy and unpredictable nature of malignant patterns. Vision Transformers (ViTs) significantly improve feature extraction that captures insights in complex images for accurate diagnosis. However, extracting local spatial in small nodules while maintaining global features is challenging due to the patch merging in the hierarchical ViTs, which is ineffective for diversified images. Thus, this work introduces a Convolutional Neural Network (CNN) and Hierarchical Vision Transformer (ViT)-assisted hybrid model for lung cancer detection, enriched by multiscale patch embedding and cross-attention fusion to improve feature extraction and analysis of lung PET/CT images. Initially, the proposed approach applies the preprocessing and augmentation procedure to improve the generalization for lung cancer detection tasks. In the hybrid model, the CNN model extracts the local spatial features from the integrated multimodal PET/CT images and divides the feature maps of images into multiple scales to provide input to the hierarchical ViT succeeded by the multiscale patch embedding and position encoding. Moreover, the design of cross-attention fusion in hierarchical ViT combines the multiscale information, allowing the model to concentrate on relevant patterns and enhance diagnostic accuracy. Thus, experimental results show that the proposed model outperforms the existing lung cancer detection approaches, particularly in cases with small or indistinct lesions, by efficiently merging multiscale embeddings.

**Keywords**: Lung Cancer Detection, Hierarchical ViT, CNN Feature Extraction, Hybrid Model, Multiscale Patch Embedding, and Cross-Attention

## INTRODUCTION

Lung cancer is the primary cause of cancer-related mortalities globally, contributing to 18% of all such fatalities [1]. The main contributor to lung cancer is the smoking habit, and its occurrence has peaked or persistently escalated in several countries, signifying that lung cancer cases may rise further in the upcoming decades. Lung cancer is a heterogeneous disease, which is primarily classified into Small-Cell Lung Carcinoma (SCLC) and non-SCLC (NSCLC). NSCLC, including adenocarcinoma, squamous cell carcinoma, and large cell carcinoma, comprise 85% of cases, whereas the remaining 15%, characterized by neuroendocrine differentiation, are SCLC cases [2]. For enhancing patient survival rates, early detection of lung cancer is performed based on lung nodule, which is a main indicator of lung cancer. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are the common imaging techniques that play a vital role in lung cancer detection [3].

In the personalized medicine era, precise lung cancer diagnosis and classification deeply rely on histological and cytological subtyping through microscopic examination using normal histochemical and additional immunohistochemical staining. These conventional methods can be time-consuming, subjective, and susceptible to errors [4]. Hence, integrating Artificial Intelligence (AI)-based tools into clinical practice has led to significant progress in the digitization of medical imaging. Recently, AI has gained traction with deep learning (DL) breakthroughs. After an initial phase of heightened expectations when concerns arose about AI possibly substituting

radiologists, the field has transformed into a practical adoption phase [5]. By training on extensive medical datasets, DL algorithms can identify patterns beyond human perception, attaining state-of-the-art performance in cancer detection and frequently outperforming human expertise. This progression improves detection accuracy and patient experience with reduced diagnostic costs by facilitating rapid and more automated diagnoses [4]. Despite the advantages of the DL models, challenges such as the requirement for large medical datasets and bias from unrepresentative training data cause distrust in their results. However, the field of DL has witnessed substantial evolutions driven by significant computational power, vast amounts of data availability, and advances in neural network architectures. Combining medical images, genomics, clinical reports, and other patient-related data allows DL algorithms to extract complex features and offer a holistic view of cancer diagnosis. Incorporating different data modalities improves detection accuracy and contributes to a more personalized and accurate approach to cancer management [6].

Convolutional Neural Networks (CNNs) have revolutionized medical imaging by learning complex data representations across several modalities, including CT, radiography, MRI, and so on [7]. However, the stationary weights and local receptive fields can limit their capability to capture distant pixel relationships, prompting research into transformer models that efficiently encode these dependencies and enhance feature representation. Transformer acts as a framework for sequence-to-sequence prediction with remarkable capability for modeling long-range sequences and outstanding results in natural language processing and machine translation. By learning correlations across all input patches using self-attention, transformers capture long-range dependencies effectively among pixels [6]. Transformer has developed as a viable substitute to CNNs, representing competitive performance across several computer vision tasks such as object detection, semantic and instance segmentation, image recognition, and image generation. Particularly, a transformer-based architecture is utilized as the detection transformer to create the first fully end-to-end target detection model. The Vision Transformer (ViT) is the first image recognition model that depends on the transformer framework [8]. ViTs have achieved state-of-the-art results in numerous vision tasks, including object detection, image classification, and video understanding.

Furthermore, ViTs show prediction errors closer to human judgment than CNNs, prompting rising interest in their adaptation for medical imaging applications to decrease the biases inherent in CNNs. During lung cancer diagnosis, vision transformers are utilized for nodule detection, tumor segmentation, cancer classification, and survival prediction [9]. Hierarchical ViTs, including swin transformers with window-based attention, MaxViT with multi-axis attention, and pyramid vision transformers with spatial reduction attention, have recently been developed to enhance performance in medical image segmentation tasks. However, these transformers' self-attention within a single attention window restricts feature processing capabilities [10].

The main contributions of the research work are outlined as follows.

- This work presents a lung cancer detection model that integrates CNN feature extraction and hierarchical ViT, enabling localized feature extraction and global context awareness enhanced by multiscale patch embedding and cross-attention.
- Modeling multiscale patch embeddings in hierarchical ViT allows capturing cancerous patterns from CNN feature maps across various spatial resolutions with position encoding, improving the ability to detect fine-grained abnormalities in lung tissue.
- Moreover, cross-attention mechanism-associated hierarchical ViT contextually integrates diverse feature scales from the feature representation of self-attention at each scale, leveraging to focus on cancer regions in the multimodal PET/CT images.
- Thus, the experimental results demonstrated improved performance against conventionalung cancer detection and classification models, providing highly accurate and early detection aids in real-time clinical decision-making.

## LITERATURE REVIEW

Radiologists encounter rising challenges, including higher workloads as well as diagnostic demands. Traditional lung cancer detection approaches need improved accuracy. Hence, recently, several studies have concentrated on diagnosing different diseases by implementing ViT to analyze medical images from different modalities. Some of them are discussed below.

### A. Improved Vision Transformer for Medical Imaging

Owing to the single-scale self-attention mechanism, the generalization ability of transformers is often limited. Hence, the Multiscale hiERarchical vIsion Transformer (MERIT) is proposed in [11] to overcome this issue by applying self-attention at multiple scales. Cascaded attention decoding (CASCADE) refines multi-stage features. Furthermore, a multi-stage feature mixing loss aggregation method named MUTATION is utilized to improve model training. However, the dependence on complex architectures and attention mechanisms will increase computational requirements, affecting its real-time application. The combination of U-shaped and transformer architectures leads to the struggle to recover spatial information during up-sampling and the loss of image features during down-sampling. An encoder-decoder architecture is proposed in [12] for medical image segmentation by combining a hybrid encoder with two expanding paths. It minimizes feature loss and enhances spatial information recovery. The hybrid encoder captures local and global pixel information. Spatial reconstruction and convergence are improved by retaining deep-supervised, independent expanding paths and the class-token sequence. Consecutive residual connections will support spatial recovery and decrease feature loss.

For improved representation of 2D and 3D medical images, a ViT-based autoencoder called ViT-AE++ is proposed in [13] that utilizes two new loss functions. One loss function is for improving self-reconstruction by capturing structured dependencies, whereas the other is for contrastive loss. Moreover, ViT-AE++ can be extended to handle 3D volumetric medical images. However, its dependence on hyperparameters like masking ratio complicates the training process and necessitates careful tuning for optimal performance. For addressing low accuracy in recognizing small or overlapping targets during image segmentation, a hybrid vision transformer with a unified-perceptual-parsing network (ViT-UperNet) is proposed in [14] that embeds self-attention in a ViT. This helps extract multi-level features and process image features hierarchically. A UperNet fuses multiscale contextual features to enhance understanding of global context and semantic information. Pre-training is performed by a masked autoencoder that strengthens visual representation and feature learning efficiency. Even though morphology is the main standard for diagnosis, substantial tools must be developed to elucidate the diagnosis. The pre-trained ViT model is proposed in [15] for classifying multiple-label lung cancer on histologic slices in both few-shot and zero-shot scenarios. The Swin transformer (Swin-S) model [16] is proposed, and its performance is evaluated in lung cancer classification and segmentation. The Swin-S model shows improved mean Intersection over Union (mIoU) in segmentation tasks. The model's accuracy is improved by pre-training. However, Swin-S failed to adapt to 3D medical images that were extensively utilized in clinical settings. As the application of self-attention in understanding temporal distances among sparse, irregularly sampled spatial features has been explored previously, two approaches are proposed in [17] for a time-distance ViT: vector embeddings of continuous-time and a temporal emphasis model for adjusting self-attention weights. However, this approach was impacted by the overrepresentation of screen-detected cancers and slow-growing lung cancer subtypes in the cohort.

A hybrid framework called HViT4Lung is proposed in [18] by combining transformers and CNNs to improve lung cancer diagnosis. Transfer learning is employed to extract features from chest CT images to detect nodules and malignancy and address challenges associated with size and location discrepancies of nodules in CAD systems. A computer-aided detection (CAD) scheme is proposed in [19] by utilizing a 3D multiscale ViT (3D-MSViT) to improve feature extraction and lung nodule prediction from 3-dimensional CT images. A local-global transformer block structure helps individually process scale patches before combining features at the global level based on the attention mechanism, thus reducing network parameters. Owing to the absence of research on evaluating the effectiveness of different optimizers for lung disease prediction within ViT models, various optimization methods like Adam, RAdam, NAdam, AdamW, Momentum, and SGDW are evaluated in [20] using a dataset with 19,003 chest X-ray images. ViT, FastViT, and CrossViT models were trained through these optimizers to compare their performance in predicting lung diseases, eventually providing strategies for enhancing ViT architectures. However, the evaluation is performed in diverse datasets with varied sample sizes.

### B. Vision Transformers in Lung Cancer Detection

The work in [21] presented a lung cancer diagnosis model by analyzing multimodal imaging data using CNNs. Enhancing pathological categorization of lung cancer types with advanced image processing techniques improves classification accuracy. Combining CNNs with Swin Transformer proposes an automatic detection scheme for lung cancer cells [22]. A mask R-CNN-based network segments microscopic images of lung cells by highlighting target cells and retaining background information using Gaussian blurring. This scheme shows reduced computation with

improved performance. It proved to be beneficial for lung cancer cell detection and classification. However, its dependence on Gaussian blur affects the robustness of the model across different imaging conditions. A lung tumor segmentation method is proposed in [23] that combines CNNs and ViTs. An encoder-decoder structure is utilized for convolutional blocks in the initial and final layers. At the same time, deeper layers integrate transformer blocks with a self-attention mechanism for comprehensive global feature mapping. However, this complex architecture causes increased training times and problems tuning hyperparameters. For evaluating the effectiveness of PET/CT images, a new approach is proposed in [24] for detecting and classifying lung cancer using DL. The detection transformer (DETR) model employs a transformer-based approach to detect tumors and support physicians in staging lung cancer patients. Segmentation needs further improvement for better tumor localization and improved tumor delineation accuracy.

Analyzing Whole Slide Images (WSIs) during histopathological examination can be error-prone and time-consuming for pathologists. Hence, the classification accuracy of histopathological images for NSCLC is improved using a DL architecture [25] that combines CNN and ViTs. ViTs analyze long-range relations between image patches, whereas CNNs capture local image features. However, its computational intensity will affect its practical deployment in resource-constrained environments. WSIs are difficult during manual pixel-wise annotation because of their high cost and large scale. Tumor heterogeneity and subtle morphological differences cause variability in expert annotations, affecting accuracy. Hence, Simple Shuffle-Remix ViT (SSRViT) [26] is proposed, which is a two-stage weakly supervised learning framework to recover discriminative tokens for creating sparse WSI representations. These representations are utilized by a transformer-based classifier called SViT to perform slide-level predictions. However, its dependence on weak labels will not fully capture the intricate details essential for accurate classification. Table 1 compares the approaches that utilize ViT for segmenting or classifying medical images.

**Table 1: Comparison of the ViT-based Existing Approaches for Medical Imaging**

| Author (year) [ref] | Architecture | Method | Dataset | Application task |
|---|---|---|---|---|
| Rahman and Marculescu (2023) [11] | Multiscale hierarchical vision transformer, cascaded attention decoding | Compute self-attention across multiple windows to enhance the model's ability to capture multiscale features | Synapse multi-organ dataset, ACDC dataset (MRI, CT) | Medical image segmentation |
| Chaoyang et al. (2024) [12] | FDR-TransUNet | Leverage class-token sequences and successive residual connections to improve accuracy and feature retention. | COVID-19 Radiography Database, e COVID-Qu-Ex Dataset (X-ray) | Medical image segmentation |
| Prabhakar et al. (2023) [13] | ViT-AE++ | Train an autoencoder for learning effective domain-specific representations of 3D volumes without labeled data | 2D chest X-ray dataset, BraTS, Erasmus Glioma Database (MRI) | Self-supervised medical image representations |
| Ruiping et al. (2024) [14] | ViT- unified-perceptual-parsing network (UperNet) | Improve long-range dependency modeling and feature fusion for small targets. | ACDC2017 (MRI) | Medical image segmentation |
| Guo and Fan et al. (2022) [15] | Pre-trained Vision Transformer | Classify multiple-label lung cancer in both Zero- and Few-Shot settings | LC25000 (histopathological images) | Lung cancer multi-label classification |
| Sun et al. (2023) [16] | Improved Swin Transformer | Employ the sliding window operation for the detection of lung cancer | LUNA16 dataset (CT) | Lung cancer image classification and segmentation |

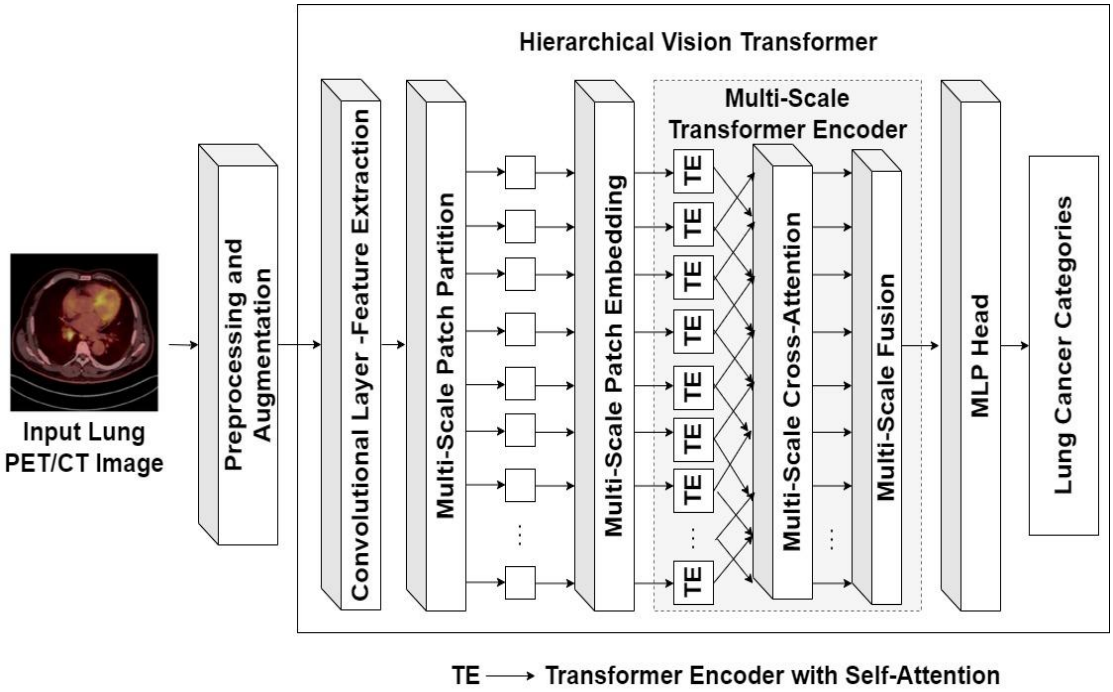
| | | | | |
|---|---|---|---|---|
| Li et al. (2023) [17] | Time-distance vision transformer | Incorporate temporal dynamics into the self-attention process | Tumor-CIFAR v1 and v2, NLST dataset (CT) | Medical imaging classification |
| Ko et al. (2024) [20] | ViT-B/16 model | Identify the most effective optimization method for training ViT | Chest X-ray dataset (chest X-ray) | Detection of lung diseases |
| Chen et al. (2022) [22] | Swin transformer | Segment and defocus background noise in cell images | Herlev and SIPaKMeD dataset (cervical cell images) | Detection and classification of lung cancer cells |
| Barbouchi et al. (2023) [24] | Transformer-based DNN | Achieve precise tumor localization and staging | Lung-PET-CT-Dx dataset (PET/CT) | Detection and classification of lung cancer |
| Imran et al. (2024) [25] | CNNs and ViT | Eliminate the need for pre- or post-processing | LC25000 dataset (histopathological images) | Non-small cell lung cancer detection and classification |

## PROBLEM STATEMENT

Lung cancer has become one of the most hazardous and increasing causes of death worldwide. Identifying small early-stage lesions in medical imaging is challenging due to delayed diagnosis increases the death rate or health risks. A lung cancer diagnosis has extensively applied conventional deep learning models, especially CNNs. Even though CNN-based models ensure accurate and timely cancer diagnosis, there is a frequent inability to grasp contextual information and intricate spatial interactions in high-resolution medical images. Traditional ViT models fail to handle the different features scales in lung cancer imaging. Hence, recent developments in ViT models have demonstrated their effectiveness in capturing global contextual information through self-attention mechanisms. For instance, lung CT scan lesions vary from small nodules to massive masses, necessitating multiscale feature representation. In hierarchical ViTs, patch merging increases speed by reducing the number of tokens; however, it ignores fine-grained information, particularly in complex tasks like medical imaging. Even though downsampling in a hierarchical model improves global context, it degrades local precision. It misses important small features affecting lung cancer detection accuracy, such as nodules in lung scan images. In addition, several multiscale patch embedding models assist in efficiently and accurately examining large medical images, but analyzing small perturbations is challenging and limits the generalization. Hence, this work aims to improve diagnostic accuracy by allowing the model to concurrently capture highly generalized contextual features and fine-grained local features in the lung PET/CT images.

## PROPOSED METHODOLOGY

The primary goal of this work is to propose Vision Transformer Model-Assisted Lung Cancer Detection with Multiscale Patch Embedding and Cross-Attention Fusion system, increasing lung cancer detection accuracy. The proposed system initially normalizes and augments lung PET/CT images in order to standardize inputs and improve the robustness of the model, which aims to capture both the global features of lung regions and the fine-grained features of possible lesions by dividing the images into patches of different sizes after extracting convolutional feature maps from the augmented images. Moreover, the cross-attention method in the hierarchical ViT enables learning contextual information across different image patches. By learning local features, such as small or tiny nodules, and global structural patterns, the proposed approach enhances lung cancer detection capabilities from the multiscale patched embeddings through cross-attention rather than patch merging. Figure 1 illustrates the proposed lung detection methodology.




### A.Medical Image Preprocessing and Augmentation

To conduct the experiment, this work employed lung images from the Cancer Imaging Archive (TCIA), namely the Lung-PET-CT-DX dataset [27], a large-scale resource designed to perform lung cancer diagnosis research. It consists of imaging data for 355 patients, with 251,135 images for 1,295 series from 436 studies. It also comprises extensive images of both CT and PET-CT images in DICOM format, as well as XML files that designate tumor locations with bounding boxes. Class labels in the Lung-PET-CT-DX dataset are based on tumor histology type, including 'A' as adenocarcinoma, 'B' as small cell carcinoma, 'E' as large cell carcinoma, and 'G' as squamous cell carcinoma tagged as lung cancer disease types for each patient.

Initially, the proposed approach applies the preprocessing procedures on the integrated lung PET/CT images, involving resizing to uniform dimensions, normalizing pixel intensities, and noise reduction to eliminate imaging artifacts. Owing to the images recorded from various CT devices, the input pixel arrays are transformed into a unified pixel range by rescaling. Subsequently, the pixel arrays of the input image are normalized into values between 0 and 1, and to further normalize the images, the resizing is also applied regarding a particular shape, 300×300×3. Moreover, data augmentation [28] improves the robustness of the model by applying flipping and rotation as the geometric transformations for the normalized images. In the image augmentation process, geometric transformation with color jittering, such as contrast, brightness, and hue adjustment, increases the diversity of the training set, thereby improving generalization for lung cancer detection tasks.

### B. Feature Extraction and Multiscale Patch Embedding

To enhance the image feature representation, the proposed approach intends to capture coarse-grained and fine-grained features from the input images by integrating the hybrid model with CNN and hierarchical ViT [29].

**CNN Feature Extraction:** In the hybrid CNN-Transformer model, the convolutional layer initially learns the augmented input image to represent the feature maps from low-level to high-level, such as shapes, edges, and textures, thereby improving the extraction of local spatial patterns. In ViT-assisted lung cancer detection, spatial feature map extraction enables the CNN to capture the hierarchical structure of input image, which is highly significant for medical image analysis, such as lung scans, when information at various scales contributes towards precise cancer detection. In the subsequence of extracting comprehensive feature maps ($F_{CL}$) by the convolution layer, the feature maps are further separated into many scales to reflect different degrees of information. For the Vision Transformer, each scale is further embedded as patches and then transformed into token sequences. Thus, the CNN enables the ViT to focus on various granularity levels within the derived feature maps, enhancing the model's capability to determine multiscale patterns relevant to cancer.

$$F_{CL} = ReLU(w_{CL} * F_{CL-1} + b_{CL}) \quad (1)$$

In equation (1), convolutional layers apply the ReLU as a non-linear activation function for the image feature extraction at layer 'L' with the convolution kernel.'$w_{CL}$' and bias '$b_{CL}$'. Thus, the CNN outcomes are a set of feature maps for the augmented input PET/CT image, which are further provided in a multiscale patch embedding and position encoding layer in hierarchical ViT.

**Multiscale Patch Embedding:** The lung PET/CT images are further divided into patches at various scales regarding feature map representation to extract these contextual features. Multiscale patch partitioning aims to examine whether larger patches capture the wider context of lung regions. In contrast, smaller patches concentrate on finer details, such as small nodules. In the patch embedding procedure, positional encodings preserve spatial relationships between patches and improve the processing ability of ViT for intricate lung structures.

In the proposed hierarchical ViT, multiscale patch embedding is a key design in capturing fine-grained and global context information from images by processing patches at multiple scales. Even though traditional hierarchical transformers achieve hierarchical representation learning through window-based multiscale patching for local and global information analysis in their early layers, patch merging in the hierarchical ViT tends to ignore the capture of fine-grained features at further layers. Even though deeper layers in the hierarchical ViTs enable the sharing of cross-scale information, ensuring global self-attention is challenging across the features with finer details due to patch merging. Hence, in contrast to traditional ViTs, the hierarchical ViTs partition the image into non-overlapping windows and apply the self-attention locally to examine the local context at different windows via the transformer encoder in each scale 'S' with the embedding ($E_{SS'}$) representation of multiscale patches. In the multiscale patch embedding, each feature map (F) is partitioned into non-overlapping patches for each image with the dimensions of height (H), width (W), and depth (D), resulting. ($[H \times W]/p^2$) patches for each image for patch size' p'.

$$E_{SS'} = w_i . Flatten(P_{SS'}) + b_i \quad (2)$$

To enhance the ability of hierarchical transformer in lung cancer detection, the proposed approach embeds the patch representation at multiple scales (S, $S'$) as mentioned in equation (2) by flattening the patches ($P_{SS'}$) in 'F' into a vector transformation with the integration of weight matrix ($w_S$) and bias ($b_S$). Thus, the proposed approach provides an embedding representation ($E_S$) for each feature map (F) at each scale to contextually extract the insights by the transformer encoder.

### C. Hierarchical Vision Transformer with Cross-Attention

In the design of hierarchical ViT, the proposed lung cancer detection model integrates cross-attention instead of relying on self-attention for multiscale image representation, enabling efficient information transmission between various feature representations and resolution levels in different scales of images. By capturing long-range relationships and local features efficiently, cross-attention significantly enhances the performance of lung cancer diagnosis through accurate fine-grained features-based lung cancer detection. In contrast to patch merging, which progressively decreases the feature map resolution, cross-attention enables patches of different scales to be embedded and processed independently while maintaining fine-grained features and structural variations at each scale.

Cross-attention allows the model to efficiently utilize both local and global context of input lung PET/CT images. As mentioned in equation (3-5), the proposed approach formulates the Query (Q), Key (K), and Value (V) for cross-attention mechanism at multiple scales, S and $S'$ at layer (i) in the transformer architecture. $W_q^S$, $W_k^{S'}$, and $W_v^{S'}$ are projection matrices for Q, K, and V, respectively, at S and $S'$ scales.

$$Q_S = X_S^i W_q^S \quad , \quad K_{S'} = X_{S'}^i W_k^{S'} \quad , \quad V_{S'} = X_{S'}^i W_v^{S'} \quad (3)$$

$$head_h^{(S,S')} = Att\left(Q_S^{(h)}, K_{S'}^{(h)}, V_{S'}^{(h)}\right) \quad (4)$$

In the multiscale embedding representation, the proposed approach computes the cross-attention score for each head (h) with the pair of ($S, S'$) scales in Q, K, V computation across 'n' number of multiple scales in which 'S' denotes a particular scale and $S'$ refers to another scale that excludes 'S'. The Multi-Head Attention (MHA) concatenates multiple heads of cross-attention with the output projection matrix.

$$X_{Fusion}^{i} = \sum_{S=1}^{n} \sum_{S' \neq S} MHA(Q_S, K_{S'}, V_{S'}) \quad (5)$$

In the proposed lung cancer detection system, cross-attention fusion ($X_{Fusion}^{i}$) enables the model to assign weight to features from each scale based on their relevance to lung cancer patterns. The selected region of the input image at each scale ensures that potential fine features, such as small malignant lesions in lung PET/CT images, receive higher attention than loss-prone feature averaging in simple patch merging. Consequently, the cross-attention process inevitably contributes to interpretability by demonstrating lung cancer-influencing features from different scales from the attention maps focused on recognizing specific regions of interest in lung cancer images. Thus, the cross-attention maps provide insights to clinicians to improve confidence and validate the effectiveness of the prediction model by highlighting potentially malignant regions in lung CTs, strengthening detection accuracy in regions with high cancer risk. Algorithm 1 describes the steps involved in the proposed lung cancer detection model.

**Input:** Lung CT Images
**Output:** Lung Cancer Categories/Types

**//Preprocessing and Augmentation//**
**for** all the input Lung Images **do**
  Apply normalization and resizing
  **for** all the preprocessed images **do**
    Apply the augmentation, including geometric transformation and color jittering
    **if** $C(A_{LI}) == C(R_{LI})$ **then**
      Generate the new training set with category-specific augmented images
    **endif**
  **endfor**
**//Feature Extraction//**
  **for** all the augmented images **do**
    Extract the local spatial features using CNN
    Represent the feature maps for each image as in Equation (1)
  **endfor**
**//Multi-Scale Patch Embedding//**
    **for** all the extracted feature maps of images **do**
      Divide each image into multiple scale patches in terms of feature maps
      **for** the multiple patches of images **do**
        Generate embedding representation with position encoding as in Equation (2)
      **endfor**
    **endfor**
**//Cross-Attention Fusion//**
    **for** the embedded representation of multi-scale patches **do**
      Apply the hierarchical ViT
      **for** each patch **do**
        Design the transformer encoder for each patch in hierarchical ViT
        Learn the local feature representation by the self-attention
      **endfor**
      Apply the cross-attention for multiple scales using Equations (3) and (4)
      Perform the Cross-attention fusion using Equation (5)
        **for** integrated feature representation with attention maps **do**
          Execute classifier head
          Categorize the lung cancer types
        **endfor**
    **endfor**
**endfor**

**Algorithm 1: Pseudocode of the Proposed Lung Detection Methodology**

## EXPERIMENTAL EVALUATION

The experimental setup for the lung cancer detection model leverages Python machine learning libraries and OpenCV to build and train the image processing algorithm. Parameters for a hybrid model designed to handle lung PET/CT images for cancer detection are detailed in Table 2.

**Table 2: Implementation Parameters**

| Parameters | | CNN-Hierarchical ViT |
|---|---|---|
| Dropout Rate | | 0.3 |
| Learning Rate | | 0.0001 |
| Activation | CNN | ReLU |
| | ViT | GELU |
| Loss Function | | Cross-Entropy |
| Epochs | | 100 |
| Batch Size | | 16 |
| Optimizer | | AdamW |

To assess the performance of the lung cancer detection model, the experimental model employs precision, recall, F1-score, and accuracy. In the lung cancer detection task, True Positives (TP) correctly identified the target lung cancer type, True Negatives (TN) identified other lung cancer types, False Positives (FP) incorrectly classified lung cancer types, and False Negatives (FN) incorrectly detected lung cancer types.

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

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

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

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

To evaluate the performance metrics on the multi-class lung image dataset, the experimental results provide the Average Precision (AP), Average Recall (AR), Average F1-Score (AF), and Average Accuracy (AA) across all the classes.

### A. Result Discussion on Histologic Classification

For the evaluation, the experimental model utilizes the patients with lung image categories of Adenocarcinoma (ADC), Squamous Cell Carcinoma (SQC), and Small Cell Carcinoma (SCC). Due to the lack of CT/PET images in the 5 Large Cell Carcinoma (LCC) category, the evaluation dataset contemplates only ADC, SCC, and SQC lung cancer types. Finally, this work exploits 1160 images belonging to 25 patients with ADC, SCC, and SQC categories.

**Table 3: Comparative Lung Cancer Detection Performance**

| Author | Dataset | Number of Test Samples | Model | AP (%) | AR (%) | AF (%) | AA (%) |
|---|---|---|---|---|---|---|---|
| Jacob, C., & Menon, G. C. (2022) [21] | Lung-PET-CT-Dx | 204 | Shallow CNN | 95.5 | 93.76 | 94.63 | 95.0 |
| Sun, R et al (2023) [16] | LUNA16 | 7076 | Swin Transformer | - | - | - | 82.3 |
| Chen, Y et al., (2022) [22] | Lung Cell | 932 | CNN + Swin Transformer | 95.2 | 92.6 | 93.88 | 96.14 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Barbouchi, K et al (2023) [24] | Lung-PET-CT-Dx | 270 | DETR Transformer | 93.0 | 96.0 | 94.0 | 96.0 |
| Imran, M et al., (2023) [25] | LC25000 | 1500 | CNN + Hierarchical ViT | 98.0 | 98.2 | 98.0 | 98.8 |
| **Proposed** | **Lung-PET-CT-Dx** | **270** | **CNN + Hierarchical ViT** | **98.92** | **98.99** | **98.95** | **99.5** |

As mentioned in Table 3, prior research in lung cancer detection [22, 25] has utilized hybrid deep learning models, and limitations remain due to insufficient representation of evolving cancer patterns in training data, which reduces the accuracy in distinguishing between various malignant regions. Furthermore, the lung cancer detection approaches [16, 22, 24, 25] have employed the transformer models to potentially extract the features, resulting in relatively minimal performance due to the lack of cross-attention fusion from the multiple scales for the complex images. In contrast, the proposed model leverages Vision Transformers enhanced with CNN-based feature extraction, multiscale patch embedding, and cross-attention fusion, enabling a deeper understanding of complex cancer features across different scales and regions. Consequently, the proposed approach enables highly complex image learning and precise tumor detection, leading to improved accuracy of 99.5% on the Lung-PET-CT-Dx dataset. Even though the work in [21] achieved comparatively higher precision and recall than the research in [24] while testing on a specific Lung-PET-CT-Dx dataset, it fails to generalize across multiple lung cancer types due to limited data diversity and inadequate analysis of complex visual patterns, proved in producing comparatively minimal accuracy as 95% with overfitting. In addition, it is confronted with large-scale and complex image patterns due to the shallow CNN architecture, which tends to ignore the analysis of global feature relationships.

Furthermore, the research works [16, 22] applied the Swin transformer, the hierarchical transformer for the lung cancer classification. However, patch merging without integrating various local spatial feature representations globally misleads the decision-making. Consequently, the work in [22] accomplished only a 93.88% F1-score while testing on the Lung cell image dataset, even though the CNN model extracts local spatial features. Also, in the comparative work [25], the combination of CNN and hierarchical ViT as similar employed in the proposed system, lack of modeling of the multiscale patch embedding in the hierarchical model, and cross-attention ignore the capturing of both the local and global features throughout the layers, resulting inaccurate detection of fine-grained lung cancer tissues in the diversified images. Thus, the proposed approach addresses these gaps, ensuring a highly adaptable, precise, and generalizable approach for lung cancer detection with a 98.95% F1 score.

### B. Evaluation on Lung-PET-CT-Dx Dataset

To assess the lung cancer detection capabilities on the Lung-PET-CT-Dx dataset, Table 4 compares the performance metrics of the proposed Vision Transformer-based model on different lung cancer types. By designing a multiscale patch embedding for the CNN feature maps and cross-attention fusion, the proposed model accomplishes higher performance on all three categories, which is comparatively higher than the existing research works [21, 24], as mentioned in Table 5. Also, the proposed approach exhibits a higher F1 score and accuracy in all three classes such as ADC, SCC, and SQC. Results in Table 4 show that the augmented samples generated by the proposed model and multiscale embedding process significantly enforce the decision-making by providing comprehensive cancer patterns for each category.

**Table 4: Proposed Lung Cancer Detection Performance on Lung-CT Dataset**

| Cancer Types | Precision (%) | Recall (%) | F1-Score (%) | Accuracy (%) |
|---|---|---|---|---|
| **ADC** | 98.81 | 100 | 99.4 | 99.58 |
| **SCC** | 99.22 | 99.12 | 99.17 | 99.73 |
| **SQC** | 98.72 | 97.86 | 98.29 | 99.19 |

The proposed system employs cross-attention-enabled hierarchical transformer architecture for PET/CT lung cancer detection, integrating image patches and spatial features to capture intricate malignancy regions. In contrast to only focusing on explicit features, the proposed approach learns the extracted patches at multiple levels of CNN feature maps that facilitate the accurate recognition of cancer regions even in complex images. Consequently, the proposed approach achieves higher accuracy at 99.58%, 99.73%, and 99.19% in the ADC, SCC, and SQC categories, respectively, demonstrating enhanced adaptability and generalization in lung cancer detection.

### C.Comparative Evaluation on Lung-PET-CT-Dx Dataset

As depicted in Table 5, the proposed Vision Transformer-based approach, incorporating Transformer and cross-attention, consistently exceeds the shallow CNN and DETR transformer in the research works [21, 24] in lung cancer detection tasks. The Vision Transformer with cross-attention excels at capturing long-term dependencies within image patches, aided by multiscale patch embedding. The cross-attention mechanism enhances the model's ability to prioritize critical lung patterns and align spatial features across different scales, significantly boosting detection accuracy. The existing work [24] is constrained by generalized feature extraction for capturing comprehensive lung features due to the limited exploration of the cross-attention of various scales while investigating the integrated multimodal PET/CT images. Because the complementary information is incorporated in the integrated PET/CT images, including metabolic activity and anatomical data, information loss is avoided significantly. To resolve this constraint, the proposed approach applies multiscale patch embedding, enabling the detection model to extract fine-grained and large features by examining the tumor region and surrounding tissue at various scales.

**Table 5: Comparative Analysis of Histologic Classification Works**

| Cancer Types | Proposed | | | Barbouchi, K et al (2023) [24] | | | Jacob, C., & Menon, G. C. (2022) [21] | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision (%) | Recall (%) | F1-Score (%) | Precision (%) | Recall (%) | F1-Score (%) | Precision (%) | Recall (%) | F1-Score (%) |
| **ADC** | **98.81** | **100** | **99.4** | 89 | 100 | 94 | 96.4 | 95.3 | 95.8 |
| **SCC** | **99.22** | **99.12** | **99.17** | 90 | 99 | 94 | 95.5 | 94.0 | 94.7 |
| **SQC** | **98.72** | **97.86** | **98.29** | 99 | 88 | 93 | 94.8 | 92.0 | 93.4 |

By leveraging the interaction between the CNN and Vision Transformer, the proposed model achieves superior feature learning and patch processing, yielding a recall of 100%, 99.12%, and 97.86% for the ADC, SCC, and SQC types, respectively. Thus, the proposed hybrid approach with multiscale patch embedding effectively adapts to dynamic imaging environments than the hybrid approach in [21] due to the information loss in the PET/CT image processing, ensuring higher precision, recall, and F1-score in lung cancer detection compared to the existing researches [21, 24] for the Lung-PET-CT-Dx dataset.

### D.Evaluation of Stages in the Proposed Lung Cancer Detection

Table 6 highlights that the proposed approach significantly enhances lung cancer detection by enhancing the ViT model with CNN-based feature extraction, multiscale patch embedding, and cross-attention fusion. This integration allows for robust feature extraction across various scales and image regions, enabling the model to detect intricate patterns in lung tissue with high accuracy. Combining CNN feature extraction to capture detailed local features with multiscale patch embedding enriches the ViT model's understanding of complex visual patterns. Cross-attention fusion further strengthens this model by aligning and integrating features from multiple patches, which improves its ability to distinguish between benign and malignant regions. This method reduces false positives and negatives, enhancing the model's precision and recall. Automated multiscale patch selection and cross-attention integration serve as key contributions to optimizing training data and providing comprehensive representations that improve detection performance.

**Table 6: Comparative Evaluation of Proposed Processes in Lung-PET-CT-Dx Dataset**

| Stages in Proposed Algorithm | Modality | Model | Lung Cancer Detection (Histologic Classification) Performance | | | |
|---|---|---|---|---|---|---|
| | | | AP (%) | AR (%) | AF (%) | AA (%) |
| Augmentation + Feature Extraction | CT | CNN | 90.19 | 89.37 | 89.78 | 90.24 |
| Augmentation + Feature Extraction + Hierarchical ViT | CT | CNN + ViT | 92.21 | 93.04 | 92.63 | 92.77 |
| Augmentation + Feature Extraction + Multiscale Patch Embedding + Hierarchical ViT | CT | CNN + ViT | 95.31 | 95.04 | 95.17 | 94.98 |
| Augmentation + Feature Extraction + Multi-scale Patch Embedding + Hierarchical ViT + Cross-Attention | CT | CNN + ViT | 96.79 | 96.81 | 96.80 | 96.08 |
| Augmentation + Feature Extraction + Multi-scale Patch Embedding + Hierarchical ViT + Cross-Attention | PET/CT | CNN + ViT | **98.92** | **98.99** | **98.95** | **99.5** |

Moreover, by inherently learning the integrated PET and CT image features within a Vision Transformer framework, the proposed approach outperforms single-modality methods for lung cancer detection, achieving higher accuracy by capturing both local and global features of lung images. This demonstrates the effectiveness of combining CNN-based feature extraction with multiscale patch embedding and cross-attention fusion, especially for detecting complex visual cues in complex lung tissue structures. In this model, CNNs enhance feature representation. At the same time, multiscale patch embedding improves spatial understanding, and cross-attention allows for refined feature alignment across scales, enabling the transformer architecture to recognize intricate patterns with minimal false positives and false negatives. The proposed approach overcomes these challenges by leveraging CNN features, patch embeddings, and cross-attention fusion, enhancing both recall and accuracy in lung cancer detection.

## CONCLUSION

This work presented an improved lung cancer detection model that combines CNN feature extraction with a Vision Transformer (ViT) architecture, which employs multiscale embedding and cross-attention. The proposed algorithm effectively captured inherent complex malignant patterns in PET/CT images by integrating CNN-based localized feature extraction with the global context capabilities of ViT. The multiscale embedding design in the hierarchical ViT significantly captured the holistic features from the integrated multimodal PET/CT images, avoiding information loss through multiscale processing. Also, the cross-attention in the hierarchical ViT comprehensively highlighted the global cancerous features from the multiple scales to detect the cancerous regions in the images precisely. The experimental results show that the proposed model outperforms the existing hybrid CNN-ViT approaches with higher detection accuracy at 99.5%, highlighting its potential to assist clinicians in diagnosing lung cancer.

## REFRENCES

[1] Thanoon, M. A., Zulkifley, M. A., Mohd Zainuri, M. A. A., & Abdani, S. R. (2023). A review of deep learning techniques for lung cancer screening and diagnosis based on CT images. *Diagnostics*, *13*(16), 2617.

[2] Davri, A., Birbas, E., Kanavos, T., Ntritsos, G., Giannakeas, N., Tzallas, A. T., & Batistatou, A. (2023). Deep learning for lung cancer diagnosis, prognosis and prediction using histological and cytological images: a systematic review. *Cancers*, *15*(15), 3981.

[3] Shariff, V. A. H. I. D. U. D. D. I. N., Chiranjeevi, P., & Km, A. (2023). An Analysis on Advances In Lung Cancer Diagnosis With Medical Imaging And Deep Learning Techniques: Challenges And Opportunities. *Journal of Theoretical and Applied Information Technology*, *101*(17), 7083-7095.

[4] Gayap, H. T., & Akhloufi, M. A. (2024). Deep machine learning for medical diagnosis, application to lung cancer detection: a review. *BioMedInformatics*, *4*(1), 236-284.

[5] Chassagnon, G., De Margerie-Mellon, C., Vakalopoulou, M., Marini, R., Hoang-Thi, T. N., Revel, M. P., & Soyer, P. (2023). Artificial intelligence in lung cancer: current applications and perspectives. *Japanese journal of radiology*, *41*(3), 235-244.

[6] Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023). Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, *85*, 102762.

[7] Fanizzi, A., Fadda, F., Comes, M. C., Bove, S., Catino, A., Di Benedetto, E., ... & Massafra, R. (2023). Comparison between vision transformers and convolutional neural networks to predict non-small lung cancer recurrence. *Scientific Reports*, *13*(1), 20605.

[8] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, *88*, 102802.

[9] Parvaiz, A., Khalid, M. A., Zafar, R., Ameer, H., Ali, M., & Fraz, M. M. (2023). Vision Transformers in medical computer vision—A contemplative retrospection. *Engineering Applications of Artificial Intelligence*, *122*, 106126.

[10] Ali, H., Mohsen, F., & Shah, Z. (2023). Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review. *BMC Medical Imaging*, *23*(1), 129.

[11] Rahman, M. M., & Marculescu, R. (2024, January). Multiscale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Medical Imaging with Deep Learning* (pp. 1526-1544). PMLR.

[12] Chaoyang, Z., Shibao, S., Wenmao, H., & Pengcheng, Z. (2024). FDR-TransUNet: A novel encoder-decoder architecture with vision transformer for improved medical image segmentation. *Computers in Biology and Medicine*, *169*, 107858.

[13] Prabhakar, C., Li, H., Yang, J., Shit, S., Wiestler, B., & Menze, B. (2024, January). ViT-AE++: improving vision transformer autoencoder for self-supervised medical image representations. In *Medical Imaging with Deep Learning* (pp. 666-679). PMLR.

[14] Ruiping, Y., Kun, L., Shaohua, X., Jian, Y., & Zhen, Z. (2024). ViT-UperNet: a hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation. *Complex & Intelligent Systems*, 1-13.

[15] Guo, F. M., & Fan, Y. (2022). Zero-shot and few-shot learning for lung cancer multi-label classification using vision transformer. *arXiv preprint arXiv:2205.15290*.

[16] Sun, R., Pang, Y., & Li, W. (2023). Efficient lung cancer image classification and segmentation algorithm based on an improved swin transformer. *Electronics*, *12*(4), 1024.

[17] Li, T. Z., Xu, K., Gao, R., Tang, Y., Lasko, T. A., Maldonado, F., ... & Landman, B. A. (2023, February). Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography. In *Proceedings of SPIE--the International Society for Optical Engineering* (Vol. 12464). NIH Public Access.

[18] Nejad, R. R., & Hooshmand, S. (2023, June). HViT4Lung: Hybrid Vision Transformers Augmented by Transfer Learning to Enhance Lung Cancer Diagnosis. In *2023 5th International Conference on Bio-engineering for Smart Technologies (BioSMART)* (pp. 1-7). IEEE.

[19] Mkindu, H., Wu, L., & Zhao, Y. (2023). 3D multiscale vision transformer for lung nodule detection in chest CT images. *Signal, Image and Video Processing*, *17*(5), 2473-2480.

[20] Ko, J., Park, S., & Woo, H. G. (2024). Optimization of vision transformer-based detection of lung diseases from chest X-ray images. *BMC Medical Informatics and Decision Making*, *24*(1), 191.

[21] Jacob, C., & Menon, G. C. (2022). Pathological categorization of lung carcinoma from multimodality images using convolutional neural networks. *International Journal of Imaging Systems and Technology*, *32*(5), 1681-1695.

[22] Chen, Y., Feng, J., Liu, J., Pang, B., Cao, D., & Li, C. (2022). Detection and classification of Lung Cancer cells using swin transformer. *Journal of Cancer Therapy*, *13*(7), 464-475.

[23] Tyagi, S., Kushnure, D. T., & Talbar, S. N. (2023). An amalgamation of vision transformer with convolutional neural network for automatic lung tumor segmentation. *Computerized Medical Imaging and Graphics*, *108*, 102258.

[24] Barbouchi, K., El Hamdi, D., Elouedi, I., Aïcha, T. B., Echi, A. K., & Slim, I. (2023). A transformer-based deep neural network for detection and classification of lung cancer via PET/CT images. *International Journal of Imaging Systems and Technology*, *33*(4), 1383-1395.

[25] Imran, M., Haq, B., Elbasi, E., Topcu, A. E., & Shao, W. (2024). Transformer Based Hierarchical Model for Non-Small Cell Lung Cancer Detection and Classification. *IEEE Access*.

[26] An, J., Wang, Y., Cai, Q., Zhao, G., Dooper, S., Litjens, G., & Gao, Z. (2024). Transformer-Based Weakly Supervised Learning for Whole Slide Lung Cancer Image Classification. *IEEE Journal of Biomedical and Health Informatics*.

[27] Lung-PET-CT-Dx Dataset, Available Online at: https://www.cancerimagingarchive.net/collection/lung-pet-ct-dx/, Accessed On November, 2024

[28] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545-563.

[29] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, *54*(10s), 1-41.

[30] Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, *56*(Suppl 3), 2917-2970.