BMC Cancer

RESEARCH

Open Access

Deep learning-based lung cancer classification of CT images

Mohammad Khalid Faizi[1*], Yan Qiang[1,5], Yangyang Wei[2], Ying Qiao[2], Juanjuan Zhao[1,3,4*], Rukhma Aftab[1] and Zia Urrehman[1]

Abstract

Lung cancer remains a leading cause of cancer-related deaths worldwide, with accurate classification of lung nodules being critical for early diagnosis. Traditional radiological methods often struggle with high false-positive rates, underscoring the need for advanced diagnostic tools. In this work, we introduce DCSwinB, a novel deep learning-based lung nodule classifier designed to improve the accuracy and efficiency of benign and malignant nodule classification in CT images. Built on the Swin-Tiny Vision Transformer (ViT), DCSwinB incorporates several key innovations: a dual-branch architecture that combines CNNs for local feature extraction and Swin Transformer for global feature extraction, and a Conv-MLP module that enhances connections between adjacent windows to capture long-range dependencies in 3D images. Pretrained on the LUNA16 and LUNA16-K datasets, which consist of annotated CT scans from thousands of patients, DCSwinB was evaluated using ten-fold cross-validation. The model demonstrated superior performance, achieving 90.96% accuracy, 90.56% recall, 89.65% specificity, and an AUC of 0.94, outperforming existing models such as ResNet50 and Swin-T. These results highlight the effectiveness of DCSwinB in enhancing feature representation while optimizing computational efficiency. By improving the accuracy and reliability of lung nodule classification, DCSwinB has the potential to assist radiologists in reducing diagnostic errors, enabling earlier intervention and improved patient outcomes.

Keywords Lung cancer computed tomography (CT), Swin transformer, Object segmentation, Object classification

Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide, with lung nodules playing a critical role in its diagnosis and management [1]. In the decade of 2020, lung cancer is expected to have caused 2.2 million new cases and 1.8 million deaths globally [2]. Lung nodules, which range in size from 3 mm to 3 cm, can be either benign or malignant. Radiologists typically assess nodule malignancy based on factors such as size, location, internal structure, and texture. However, small nodules ($< 3mm$) are often overlooked due to their asymptomatic nature and limited visibility on CT scans. The current diagnostic process heavily depends on the expertise of radiologists, but the growing population and the decreasing availability of qualified healthcare professionals have shifted focus from early intervention to diagnostic medicine. Studies have shown that radiologists without Computer-Aided Diagnosis (CAD) tools exhibit high false-positive rates, ranging between 51% and 83.2%, while their sensitivity is between 94.4% and 96.4%.

*Correspondence:
Mohammad Khalid Faizi
khalidfaizi840@gmail.com
Juanjuan Zhao
zhaojuanjuan@tyut.edu.cn
[1] College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Taiyuan 030024, Shanxi, China
[2] First Hospital of Shanxi Medical University, Taiyuan 030001, Shanxi, China
[3] School of Software, Taiyuan University of Technology, Taiyuan, Shanxi, China
[4] College of Information, Jinzhong College of Information, Jinzhong, Shanxi, China
[5] School of Software, North University of China, Taiyuan, Shanxi, China

Over the years, CAD systems have evolved from traditional feature extraction techniques such as SIFT, HOG, and LBP to advanced deep learning methods, specifically Convolutional Neural Networks (CNNs), which have become the gold standard in medical image analysis. Recent approaches also incorporate Vision Transformers (ViT), such as TransUNet [3], which combine local feature extraction with global dependency modeling, offering promising results for medical image segmentation tasks. Additionally, models like FiT [4], GVT [5], and SpikingResformer [6] further refine the model's capability by improving computational efficiency and energy usage, while enhancing feature extraction.

Despite the advancements in deep learning models, Vision Transformers (ViTs) in medical imaging still face significant limitations. They often struggle to simultaneously capture both local features, which are crucial for detecting small nodules, and global features, which provide the contextual understanding necessary for accurate classification of benign and malignant cases. Additionally, these models are computationally expensive, limiting their practicality in clinical environments where real-time processing and efficiency are essential. Addressing these issues forms the foundation of the DCSwinB model, which combines the strengths of CNNs and ViTs with enhanced local feature extraction through the Conv-MLP module and a dual-branch architecture, improving both classification accuracy and computational efficiency.

DCSwinB combines the power of the Swin-Tiny Vision Transformer (ViT) with significant modifications to enhance feature extraction efficiency. The architecture includes two parallel branches: one for local feature extraction using CNNs, and the other for global feature extraction using the Swin Transformer. This dual-branch design allows the model to capture both low-level and high-level features from CT images, improving classification accuracy while optimizing computational efficiency.

A key innovation in DCSwinB is the integration of a Conv-MLP module within the Swin Transformer branch. This module strengthens the connections between adjacent windows, allowing the model to better capture long-range dependencies-a crucial capability for analyzing high-dimensional medical images like CT scans. The hierarchical structure of DCSwinB consists of four stages, with 2, 2, 6, and 2 layers of Swin Transformer building blocks, respectively. Feature maps are downsampled between stages through a patch merging procedure, and positional embeddings are incorporated to capture spatial information.

Pretrained on the LUNA16 and LUNA16-K datasets, DCSwinB outperforms existing models in classifying benign and malignant pulmonary nodules through ten-fold cross-validation. The integration of Conv-MLP within the Swin Transformer allows the model to efficiently process both local and global features, achieving state-of-the-art results while maintaining computational efficiency, making it suitable for clinical applications.

1. A dual-branch Vision Transformer architecture that extracts semantic context from CT images for improved lung nodule classification.
2. The Conv-MLP module that enhances local feature extraction and enables the capture of long-range dependencies across 3D CT images.
3. Pretraining on LUNA16 and LUNA16-K datasets followed by ten-fold cross-validation showing significant improvements in classifying lung nodules as benign or malignant.

Through this work, we aim to bridge the gap between traditional CAD methods and modern deep learning techniques, particularly Vision Transformers, to enhance the diagnosis of lung cancer and related diseases in clinical practice.

Related work

The application of deep learning techniques has revolutionized target detection, particularly within the domain of medical imaging. Various architectures have been developed, each with distinct advantages and limitations, primarily focusing on Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and hybrid approaches combining both.

CNNs [7] have been foundational in advancing medical image analysis. Their architecture, characterized by weight sharing and local connections, significantly reduces model complexity and training duration compared to traditional fully connected networks. CNNs excel at preserving the spatial hierarchy of features within an image through their convolutional layers and pooling operations. They exhibit robustness against variations such as translation, rotation, and scale changes, which is crucial for analyzing medical images where targets can appear at different positions, orientations, and sizes. The convolutional filters are adept at capturing local patterns and textures, essential for identifying subtle indicators like lung nodules in CT scans [8, 9]. In tasks like lung nodule identification, CNNs extract these local features, which are then processed by subsequent layers to predict nodule location and classification [8, 9]. A primary limitation of standard CNNs is their restricted receptive field, making it challenging to capture long-range dependencies across distant parts of an image. This is a significant drawback when analyzing complex, high-dimensional data like volumetric CT scans where global context is often

important for accurate diagnosis. Within CNN-based object detection, methodologies are often categorized as anchor-based or anchor-free. Anchor-based methods (e.g., R-CNN [10], Fast R-CNN [11], Faster R-CNN [12]) utilize predefined anchor boxes to guide localization. Single-stage detectors (e.g., YOLO [13], SSD [14], RetinaNet [15]) perform classification and localization directly. Anchor-free methods (e.g., CenterNet [16], CornerNet [17]) predict key points like object centers or corners. While successful in general computer vision, the application of many of these specific models to lung nodule classification has been somewhat limited, partly due to the aforementioned difficulty in efficiently handling long-range dependencies inherent in CT data.

The limitations of CNNs in capturing global context led to the adoption of Vision Transformers (ViTs) [18] in medical imaging. Inspired by the success of transformers in natural language processing, ViTs apply the self-attention mechanism directly to sequences of image patches, enabling them to model relationships across the entire image. ViTs excel at capturing long-range dependencies by considering the relationships between all pairs of image patches via the self-attention mechanism. This allows them to understand the global context of an image, which can be crucial for interpreting complex scenes or large structures in medical scans. The transformer architecture is inherently flexible. Variants like hierarchical ViTs (e.g., Swin Transformer [19]) improve scalability and efficiency by processing images at multiple resolutions and using shifted windows for attention calculation. Swin Transformers incorporate patch merging and Relative Position Bias (RPB) to handle large token sequences more effectively [19]. Other adaptations like FiT [4] offer adjustable patch sizes for better computational efficiency in multi-resolution scenarios. Standard ViTs process images by dividing them into fixed-size patches. This patching process can sometimes disrupt fine-grained local details and spatial contiguity, which might be less effective for tasks requiring precise localization or analysis of small objects, such as identifying small lung nodules. ViTs typically require large datasets for effective training compared to CNNs, as they lack the inductive biases (like locality and translation equivariance) inherent in convolutional operations. Furthermore, the self-attention mechanism has a quadratic complexity with respect to the number of patches, leading to significant computational overhead, especially for high-resolution images common in medical imaging. Numerous efforts have focused on enhancing ViT efficiency and effectiveness, exploring aspects like frequency domain processing [20], integration with state-space models like Mamba [21–25], improved training strategies [26], token processing [27], and architectural variations [28, 29].

Recognizing the complementary strengths of CNNs and ViTs, hybrid architectures have emerged as a powerful approach, particularly in medical imaging where both fine-grained local details and broader global context are often critical. These models aim to leverage the local feature extraction prowess of CNNs and the long-range dependency modeling capabilities of transformers within a single framework. Hybrid models effectively combine the strengths of both worlds - CNNs capture intricate local patterns and textures, while transformers model global relationships and context across the image. This synergy often leads to richer and more comprehensive feature representations. By integrating both local and global perspectives, models like TransUNet [3] (which combines a U-Net CNN encoder with a transformer) have demonstrated improved accuracy and robustness in tasks like medical image segmentation compared to using either architecture alone. The hybrid design allows for specific adaptations. For instance, E-TransUNet [30] incorporates Res2 Net modules to boost CNN feature extraction, while TransUNetRT [31] focuses on optimizing runtime. Other innovations like GVT [5] and SpikingResformer [6] aim to reduce the computational cost associated with transformers within the hybrid structure, enhancing efficiency. Some hybrid models are tailored to specific medical imaging challenges, such as GLoG-CSUnet [32] which targets noise reduction and boundary delineation in CT scans. Designing an optimal hybrid architecture involves carefully balancing the contributions of the CNN and transformer components. Achieving the right balance between local detail and global context without one overshadowing the other can be challenging. Hybrid models can be more complex to design, implement, and tune compared to single-architecture models. While some hybrid designs aim for efficiency, combining two potentially complex architectures can still lead to significant computational demands, which remains a concern for deployment in resource-constrained clinical environments.

Significant progress has been made in areas like breast cancer detection using mammograms, employing techniques ranging from dual-view models and ensemble approaches to generative methods for data augmentation [33–36] Furthermore, beyond the detection and classification of nodules, another critical challenge in oncology is the accurate subtyping of cancers, which significantly impacts treatment decisions and patient prognosis. Recent research has explored deep learning methods for classifying lung cancer subtypes directly from CT images, sometimes augmented with synthetic pathological priors

[37]. Other approaches focus on developing task-specific embeddings for few-shot classification of cancer molecular subtypes [38] or using PET/CT data with deep learning to non-invasively discriminate pathological subtypes of non-small cell lung cancer [39]. Addressing these complex tasks often requires sophisticated models capable of integrating diverse data types and capturing subtle differentiating features, highlighting the continuous evolution of deep learning in medical diagnostics.

Despite these challenges, the pursuit of effective hybrid models continues, as they offer a promising direction for developing models that capture both local and global information efficiently. The DCSwinB model proposed in this work falls into this category, aiming to provide an optimized dual-branch structure for lung nodule classification.

Proposed method

The following section will present a comprehensive conceptual overview of the proposed DCSwinB framework and its components. The subject matter will commence with an overview of the comprehensive network design, followed by an examination of the presented Dual CNN Swin Transformer Module DCSwinB.

Overall structure of model

The DCSwinB model is introduced as an innovative and efficient image classifier for extracting semantic context information from CT images. The framework of DCSwinB is illustrated in Fig. 1. The DCSwinB method is employed to investigate longrange dependencies in HRRS images in our DCSwinB structure, with the hierarchical ViT method (Swin-Tiny [27]) serving as the baseline. Then, the modelś capacity to extract longrange dependencies from 3*D* images is improved by incorporating Conv-MLP [33] into each stage of the DCSwinB. This is achieved by fortifying the ViT adjacent window connections. Finally, a dual-branch structure referred to as "CNNs + Swin Transformers" has been developed to separate every level characteristic in the initially developed Swin-Tiny into two distinct branches. This architecture not only amalgamates the advantages of a CNN section for specific characteristics and a ViT expand for general characteristics, but also realizes a portable version by mitigating the computational demands of select fully linked layers and the multiheaded attention mechanism in ViT.

The DCSwinB method employs a hierarchical vision transformer constructed by a sequence of four successive phases. The four phases comprise 2, 2, 6, and 2 layers of Swin Transformer structures, respectively. Additionally, feature maps are downsampled between each stage, with the exception of the final stage, using a patch merging procedure that is based on linear layers.

Given an 3*D* image $x \in R^{H \times W \times C}$, where H denotes the image height, W denotes the image width, and C denotes the number of image channels. DCSwinB initially applies a mask of size $n \times n$ to the input x in the convolution layer, resulting in downsampling of the 3*D* image to the first stage. This process generates a collection of

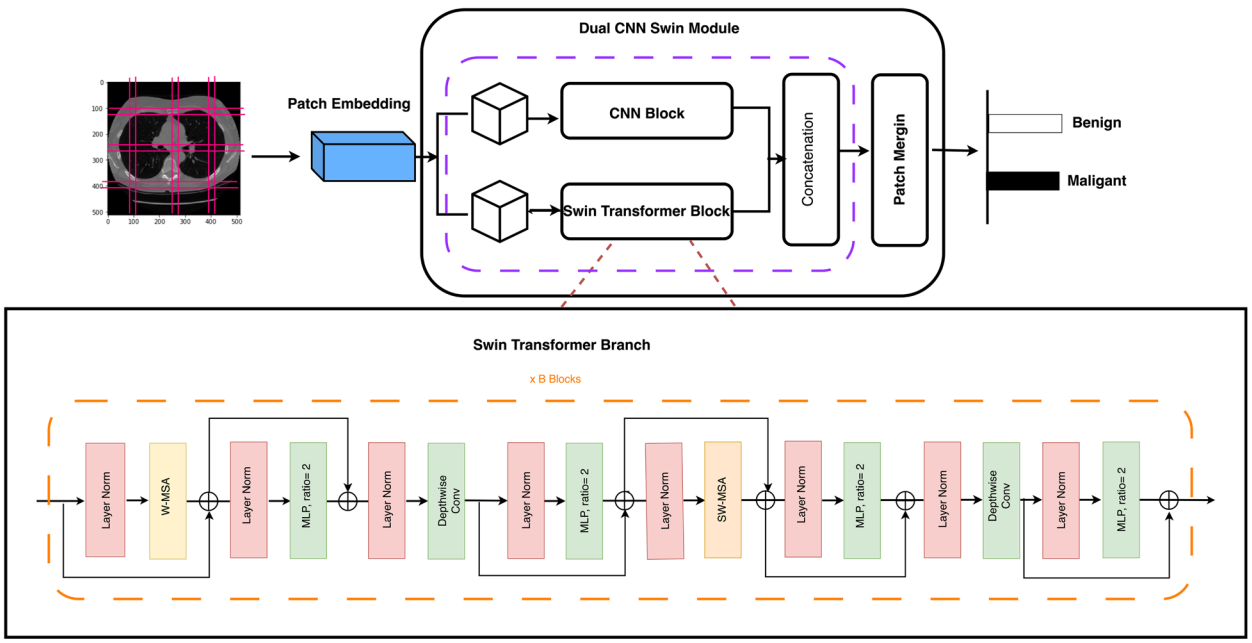

Fig. 1 overall framework of Dual CNN Swin Module

patches $P = \{p_1, p_2, \cdots, p_{n^2}\}$. Then, linear embedding is employed to tokenize each region in *P*. Furthermore, positional information is represented by incorporating relative position embeddings into these tokens. Lastly, the vector that embeds the sequence is supplied through all 4 cascaded steps of the DCSwinB, and the output elements are downsampled at each stage to *H*. Specifically, we assigned the value of 4 to *n* and the value of 96 to *C*.

Dual block CNN with SwinTransformer

To simplify, let's assume that the input for a specific stage of the hierarchical DCSwinB technique is denoted as z^{l-1}. In this method, the dual-branch structure of DCSwinB divides it into two parts: z_1^{l-1} and z_2^{l-1}. This division is achieved using a 1 x 1 convolutional layer. Additionally, the number of channels for both features is set to be half of z^{l-1}. The computational procedure can be articulated in the following manner:

$$\begin{aligned} z_1^{l-1} &= Conv_{1\times1}\left(z^{l-1}\right) \\ z_2^{l-1} &= Conv_{1\times1}\left(z^{l-1}\right) \end{aligned} \tag{1}$$

The window structuring method is used to calculate the output z_1^l of the l_layer in the transformer encoder for z_1^{l-1}. Initially, the transformer encoder is modified to include greater multilayer perceptrons (MLP) in order to improve the ViT branch's ability to capture long-range dependency information.Additionally, aiming to increase the connectivity among neighboring windows, we present a depthwise convolution between two MLP blocks, drawing inspiration from Conv-MLP, in response to the likely limitations of extra MLP layers in the ViT branch, which have a constraints on the spatial interaction information. The calculation procedure can be identified as

$$\begin{aligned} z_1^{\cdot l} &= \text{W\_MSA}\left(\text{LN}\left(z_1^{l-1}\right)\right) + z_1^{l-1} \\ z_1^{\cdot\cdot l} &= \text{MLP}\left(\text{LN}\left(z_1^{\cdot l}\right)\right) + z_1^{\cdot l} \\ z_1^{\cdot\cdot\cdot l} &= \text{DW\_Conv}\left(\text{LN}\left(\left(z_1^{\cdot\cdot l}\right)\right)^T\right) \\ z_1^{\cdot\cdot\cdot l} &= \text{MLP}\left(\text{LN}\left(\left(z_1^{\cdot l}\right)\right)^T\right) + z_1^{\cdot\cdot l} \end{aligned} \tag{2}$$

W_MSA denotes window-based multihead selfattention, LN suggests LayerNorm, MLP implies multilayer perceptron, *DW_Conv* represents depthwise convolution, and *T* represents the transpose matrix. *DW_Conv* is introduced between two MLPs, performing as a 3×3 convolution layer having the identical channel as the two MLPs. This results in an expansion of the connections within the neighboring window.

The shifted window splitting strategy is subsequently employed to calculate the resulting value z_1^{l+1} of the $l+1$ layer in the transformer encoder. The related outcome for the swin transformer branch is formed by.

$$\begin{aligned} z_1^{\cdot l+1} &= \text{SW\_MSA}\left(\text{LN}\left(z_1^{l}\right)\right) + z_1^{l} \\ z_1^{\cdot\cdot l+1} &= \text{MLP}\left(\text{LN}\left(z_1^{\cdot l+1}\right)\right) + z_1^{\cdot l+1} \\ z_1^{\cdot\cdot\cdot l+1} &= \text{DW\_Conv}\left(\text{LN}\left(\left(z_1^{\cdot\cdot l+1}\right)\right)^T\right) \\ z_1^{\cdot\cdot\cdot l+1} &= \text{MLP}\left(\text{LN}\left(\left(z_1^{\cdot\cdot\cdot l+1}\right)\right)^T\right) + z_1^{\cdot\cdot\cdot l+1} \end{aligned} \tag{3}$$

SW_MSA refers to the shifting window-based multihead selfattention. The number of parameters is reduced by setting all MLP a prolongation layers in the ViT branches to 2. Formulas (2) and (3) in the ViT branch involve intricate computational procedures. However, the initial input z_1^{l+1} for these formulas is only half the size of the original input feature z^{l-1} for each stage of the DCSwinB. The suggested dual-branch structure allows the model to acquire effective feature representation and reduces computational complexity, resulting in a portable model.

Subsequently, the CNN block initially implements a 3×3 convolution layer to extract features for the output z_2^{l-1} of Formula (1). In order to preserve robust features and expedite model convergence, max pooling is implemented, resulting in the subsequent output z_2^{l+1}.

$$z_2^{l+1} = \text{Maxpool}\left(\text{Conv}_{3\times3}\left(z_2^{l-1}\right)\right) \tag{4}$$

The output z^{l+1} of the dual-block Combining CNNs Swin Transformer module is obtained by concatenating z_1^{l+1} and z_2^{l+1} in the channel dimension. This can be expressed as:

$$z^{l+1} = \text{Concat}\left(z_1^{l+1} z_2^{l+1}\right) \tag{5}$$

The dual-branch module splits the input feature z^{l-1} into two parts, z_1^{l-1} and z_2^{l-1}, based on the channel dimension. The ViT branch of the incorporated Conv_MLP utilizes the z_1^{l-1}part to improve connections between neighboring windows while enhancing global information understanding. For enhanced global feature and model simplicity, a convolution and max pooling layer are added to the z_2^{l-1} component. Using only a convolutional layer and max pooling layer, the z_2^{l-1} component simplifies multihead selfattention and MLP processing. Consequently, the DCSwinB technique experiences a substantial decrease in computation and parameter count compared to the baseline SwinTiny approach.

The original architecture's parameters are employed to conduct the regression of classification classes and bounding boxes. (Lin et al., 2017). In this experiment, the lung nodule is being discovered and classified based on its malignancy or texture.

Depthwise convolution in Conv-MLP for efficient local interaction

To improve the ViT branch's capability to model both global dependencies and localized spatial context, we extend the standard Transformer architecture by integrating a depthwise convolution denoted as DW_Conv between two multilayer perceptrons (MLPs), as described in Equations (2) and (3). This configuration forms the Conv-MLP block, inspired by lightweight CNN-Transformer hybrids. Given the intermediate output $z_1^{\cdots l} \in \mathbb{R}^{H\times W\times \frac{C}{2}}$ after the first MLP, we apply LayerNorm and transpose it to $z_1^{\cdots l^T} \in \mathbb{R}^{\frac{C}{2}\times H\times W}$, preparing it for convolutional operation in the spatial domain. The depthwise convolution applies a distinct 3×3 kernel to each channel independently:

$$z_1^{\cdots l+1} = \mathrm{DW\_Conv}\left(\mathrm{LN}\left(\left(z_1^{\cdots l+1}\right)\right)^T\right)$$

This is followed by another MLP and residual connection, completing the Conv-MLP block:

$$z_1^{\cdots l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\left(z_1^{\cdots l+1}\right)\right)^T\right) + z_1^{\cdots l+1}$$

The depthwise convolution increases spatial connectivity within each Swin window while preserving the computational benefits of the window-based attention mechanism. Since it operates independently per channel, the parameter complexity is significantly reduced-from $3 \times 3 \times C \times C$ (in standard convolution) to $3 \times 3 \times C$ in the depthwise case.

Furthermore, as the ViT branch only processes $\frac{C}{2}$ channels (i.e., $z_1^{l-1} \in \mathbb{R}^{H\times W\times \frac{C}{2}}$), this design further lowers computational cost while retaining semantic richness. This integration balances local spatial modeling and global attention, contributing to the lightweight and efficient nature of the DCSwinB module.

Experiment and discussion

Dataset

This research utilized the publicly available LUNA16 dataset [40], which is derived from the Lung Image Database Consortium image collection (LIDC-IDRI) [41]. The LIDC-IDRI dataset comprises thoracic computed tomography (CT) scans annotated by multiple radiologists, making it a standard benchmark for pulmonary nodule analysis.

Consistent with the LUNA16 methodology, specific inclusion and exclusion criteria were applied to the LIDC-IDRI scans. CT scans exhibiting inconsistent slice spacing, missing slices, or a slice thickness greater than 3 mm were excluded to ensure data quality and uniformity. Furthermore, nodules with a diameter less than 3 mm were filtered out, as these are often considered clinically insignificant and challenging to reliably annotate.

All CT scans were resampled to a uniform isotropic voxel spacing of $1 \times 1 \times 1$ mm using trilinear interpolation. This step standardizes the spatial resolution across different scans, ensuring consistent nodule size and shape representation. The raw voxel intensities, represented in Hounsfield Units (HU), were first clipped to a typical lung window range of $[-1000, 400]$ HU to focus on relevant tissue densities. Subsequently, these clipped values were normalized to a floating-point range of [0, 1] using min-max scaling based on the specified window. For each annotated nodule, a Region of Interest (ROI) cube of size $64 \times 64 \times 64$ voxels, centered on the nodule's centroid coordinates provided in the LUNA16 annotations, was extracted. This focuses the model's input on the nodule and its immediate surroundings.

The LUNA16 dataset provides malignancy annotations based on the consensus of four experienced radiologists who graded nodules on a scale from 1 (highly unlikely to be malignant) to 5 (highly suspicious of malignancy). Following common practice, a binary classification scheme was adopted: nodules with an average radiologist rating of 3 or higher were labeled as malignant, while those with an average rating below 3 were labeled as benign. While the LIDC-IDRI dataset is known for inherent inter-reader variability, using the average score and the defined threshold helps establish a consistent ground truth for this binary task, although some level of label noise may persist.

To enhance the model's robustness to variations in nodule appearance and position, and to mitigate overfitting, extensive data augmentation techniques were applied to the extracted ROIs during training. These included: Random rotations around each axis within a range of +/- 15 degrees. Random scaling by a factor between 0.9 and 1.1. Random translations along each axis by up to $+/-5$ voxels. Random flipping along the horizontal and vertical axes with a probability of 0.5 for each.

To ensure robust evaluation and prevent data leakage between training and testing phases, the dataset was split at the patient level. This means all ROIs belonging to a single patient were assigned exclusively to either the training, validation, or test set. An 80:10:10 split ratio was used, allocating 80% of patients for training, 10% for validation (model tuning), and 10% for final testing. Furthermore, a 10-fold cross-validation strategy was employed,

also performed with patient-level splits. Within each fold, stratified sampling was used to ensure that the distribution of benign and malignant nodules was approximately maintained across the training and validation subsets for that fold. The final performance metrics reported are the average results across all 10 folds on their respective test sets, providing a reliable and unbiased estimate of the model's generalization capability.

Implementation detail

The experiment was conducted using an Anaconda environment on an Ubuntu 22.04 system with Python, leveraging the PyTorch framework [42] and CUDA for GPU-accelerated training. Hyperparameters included a learning rate of 0.001, dropout rate of 0.5 for regularization, L2 regularization with a weight decay of 0.0001, and a batch size of 32. The model was trained on an NVIDIA RTX 4060 GPU with 16 GB of memory, which enabled efficient training, completing the full training in approximately 12 hours using 10-fold cross-validation.

Hyperparameters

In the training procedure, any nodules with an ambiguity score of 3, indicating uncertainty between benign and malignant classification, were excluded to ensure clear and definitive labeling. Benign nodules were classified with degrees 1 and 2, while malignant nodules were assigned degrees 4 and 5. This resulted in a dataset of 450 malignant nodules and 554 benign nodules, drawn from a total of 1004 nodules in the LUNA16 database. The image data was then randomly partitioned into 10 subgroups for ten-fold cross-validation, with each fold using 9 subsets for training and 1 subset for testing.

Table 1 Hyperparameters configuration

Parameters	Batch Size	Epoch	Learning Rate	Weight Decay
Values	8	300	0.01	0.0001

The learning process was optimized using the Stochastic Gradient Descent (SGD) algorithm with a momentum value of 0.9 to speed up convergence and help the model escape local minima. To prevent overfitting, dropout regularization was applied. The model was trained over 300 epochs, with an initial learning rate of 0.01, which was gradually reduced during the training process to improve convergence. Specifically, the learning rate was reduced to 0.001 after 60 epochs, and further decreased to 0.0001 after 120 epochs. This learning rate decay strategy is essential for balancing fast learning during the initial phase and precise fine-tuning near the optimal solution to avoid overshooting. A weight decay of 0.0001 was applied to regularize the weights and further prevent overfitting. Network performance in Receiver Operating Characteristics of DCSwinB is shown in Fig. 2.

The batch size of 8 was chosen to strike a balance between efficient GPU usage and stable gradient estimation. A smaller batch size helps prevent the model from being biased by outliers and provides a smoother training process by introducing more randomness into

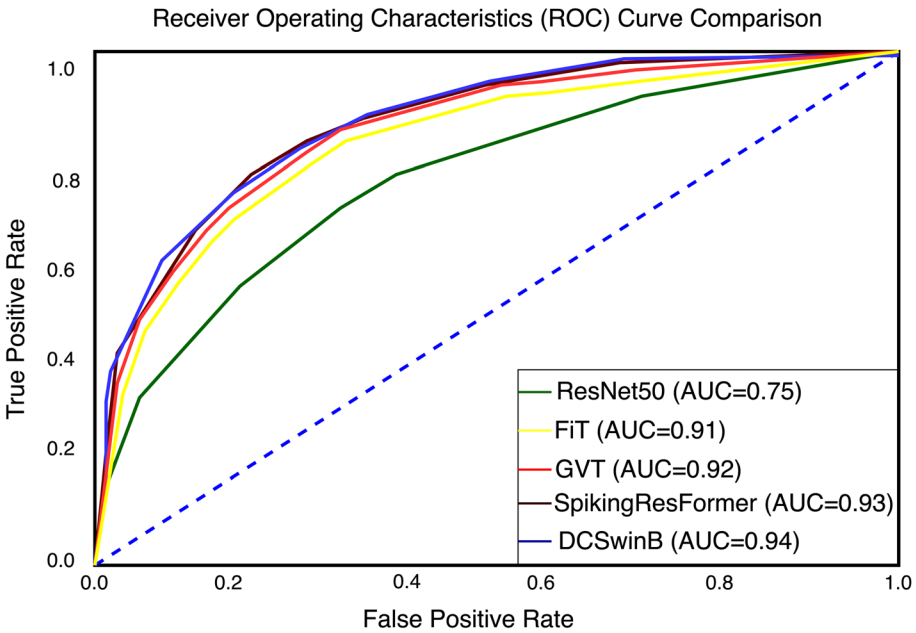

Fig. 2 Network performance of the DCSwinB is measured using the ROC

the gradient updates, which can improve generalization. However, it was chosen to be small enough to avoid excessive computational burden, especially on memory-limited systems, while still benefiting from the regularization effect of small batches. Hyperparameters are shown in Table 1

Comparison result

We compare our proposed approach with recent concurrent studies on vision transformers. The performance of the proposed model, DCSwinB, was evaluated using ten-fold cross-validation, and the results were computed as the mean performance across multiple evaluation metrics, including Accuracy, Recall,, Specificity, AUC, Precision, and F1-score. These metrics provide a comprehensive view of the model's ability to classify benign and malignant pulmonary nodules. The DCSwinB network was compared to several other deep learning models, including traditional CNN based model VGG16, ResNet50, DenseNet, Transformer-Based models Swin-T, ConvNeXt, DaViT, and CrossViT and advanced Hybrid models SpikingResformer, GVT and FiT such as which are also employed for pulmonary nodule classification tasks. Table 2 displays the data obtained from the experiment.

As shown in Table 2, DCSwinB consistently outperforms all other models, including traditional CNN-based models (VGG16, ResNet50, DenseNet) and advanced Transformer-based architectures (Swin-T, ConvNeXt, DaViT, CrossViT). Furthermore, DCSwinB surpasses recent hybrid models such as FiT, GVT, and SpikingResformer. Specifically, DCSwinB achieves the highest accuracy of 87.94%, exceeding SpikingResformer (86.96%) by 0.98%. In terms of recall, crucial for detecting malignant nodules, DCSwinB attains 85.56%, improving upon SpikingResformer (84.96%) and FiT (82.31%). DCSwinB also records the highest AUC (0.94), confirming its superior capability in both benign and malignant nodule classification.

Table 2 Performance comparison of various models on the LUNA16 dataset for lung nodule classification

Model	Accuracy (%)	Recall (%)	Specificity (%)	AUC	Precision (%)	F1-score (%)
VGG16	81.35	70.64	70.52	0.70	69.64	70.64
ResNet50	82.36	72.53	72.88	0.75	71.64	72.53
DenseNet	82.58	75.96	75.31	0.77	75.43	75.96
Swin-T	83.35	76.02	76.55	0.76	76.64	76.02
ConvNeXt	83.58	78.96	78.45	0.87	78.64	78.96
DaViT	84.58	79.96	79.80	0.88	79.64	79.96
CrossViT	84.58	80.96	80.95	0.89	80.64	80.95
FiT	85.24	82.31	84.20	0.91	82.64	82.30
GVT	85.66	83.55	84.45	0.92	83.64	83.55
SpikingResformer	86.96	84.96	84.90	0.93	84.64	84.96
DCSwinB	**87.94**	**85.56**	**85.65**	**0.94**	**85.56**	**85.56**

Bold values represent the best performance. Recall and F1-score are equivalent when Precision equals Recall

Table 3 Performance comparison of various models on the LUNA16-K dataset for lung nodule classification

Model	Accuracy (%)	Recall (%)	Specificity (%)	AUC	Precision (%)	F1-score (%)
VGG16	82.35	81.64	79.52	0.79	79.64	81.64
ResNet50	83.36	81.96	80.88	0.80	80.34	81.96
DenseNet	84.00	81.96	81.31	0.81	81.43	81.96
Swin-T	84.35	84.02	85.55	0.84	85.51	84.02
ConvNeXt	85.58	84.96	85.45	0.85	85.85	84.96
DaViT	86.58	85.36	86.80	0.85	86.34	85.36
CrossViT	87.58	85.96	86.95	0.86	86.54	85.96
FiT	89.24	86.31	88.20	0.88	88.14	86.31
GVT	89.66	86.55	88.45	0.88	88.34	86.55
SpikingResformer	89.96	88.96	89.19	0.89	89.64	88.96
DCSwinB	**90.96**	**90.56**	**89.65**	**0.90**	**85.56**	**90.56**

Bold values represent the best performance. Recall and F1-score are equivalent when Precision equals Recall

On the LUNA16-K dataset (Table 3), DCSwinB again establishes itself as the best-performing model. It achieves an accuracy of 90.96%, improving over SpikingResformer (89.96%) by 1.00%. DCSwinB also records the highest recall (90.56%), surpassing SpikingResformer by 1.60%. The F1-score of 90.56% further demonstrates balanced precision and recall, ensuring robustness in classification. Compared to ResNet50, DCSwinB shows a remarkable 4.60% improvement in recall (90.56% vs. 85.96%), a critical advantage for early detection of malignant cases. Moreover, DCSwinB achieves the highest specificity (89.65%), reducing false positives and ensuring reliable classification of benign nodules.

Although Swin-T performed competitively with an AUC of 0.90, it lags behind DCSwinB in both recall and accuracy. Overall, DCSwinB demonstrates a significant and consistent improvement in classification performance across both datasets, validating its effectiveness and robustness for clinical lung nodule diagnosis.

Ablation study

This research conducts ablation tests to further confirm the efficacy of the Dual Branch CNN-SwinT approach. The demonstration models were trained from the beginning to ensure a fair comparison. The training ratios for the LUNA16, LUNA16-K Dataset, were set at 80%, 10%, and 10% correspondingly. The four models - Swin-T, Swin-DMLP, Swin-DB, and the proposed CNN-SwinT- are depicted with their respective training accuracy curves shown in Fig. 3 on the LUNA16 validation set and Fig. 4 on the the LUNA16-K validation set. The precision of all of them exhibited rapid improvement over the first 140 epochs, succeeded by a progressive enhancement from epochs 140 to 260, ultimately culminating in a stable convergence between epochs 260 and 300.

As shown in Table 4, the proposed DCSwinB model with Conv-MLP connections achieves significant performance gains compared to the Swin-Tiny baseline. Specifically, DCSwinB reaches an accuracy of 90.96%, outperforming Swin-Tiny (87.94%) by 3.02%. The recall, which reflects the model's sensitivity to malignant nodules, improves from 85.56% to 90.56%, highlighting DCSwinB's enhanced ability to detect critical cases crucial for early lung cancer diagnosis. Moreover, DCSwinB achieves a higher specificity (89.65%) compared to Swin-Tiny (85.65%), indicating a stronger capability in correctly identifying benign nodules and reducing false positives. The area under the ROC curve (AUC) also improves from 0.92 (Swin-Tiny) to 0.94 (DCSwinB), further demonstrating superior classification robustness. In addition to performance gains, the introduction of Conv-MLP connections enhances local feature interactions through depthwise convolutions while maintaining global feature modeling via the Swin Transformer. Although not explicitly shown in parameter counts in this table, DCSwinB achieves better training efficiency, faster convergence, and stronger discriminative power - particularly for challenging cases - due to the efficient processing of local and global contexts. Overall, the combination of local

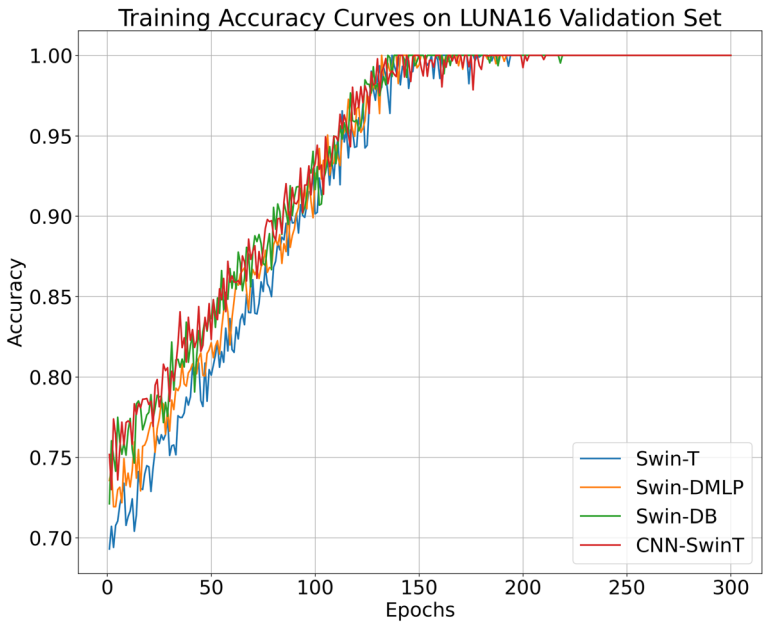

Fig. 3 Training accuracy curves on LUNA16 validation is measured using the ROC

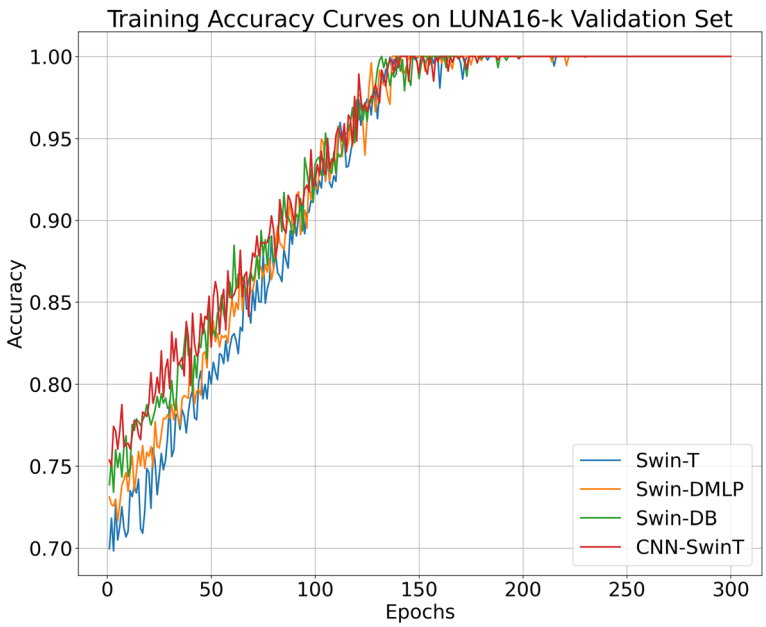

Fig. 4 Training accuracy curves on LUNA16-K validation is measured using the ROC

Table 4 Performance comparison of DCSwinB with and without Conv-MLP connections on the LUNA16-K dataset

Model	Accuracy (%)	Recall (%)	Specificity (%)	AUC	Precision (%)	F1-score (%)
Swin-Tiny	87.94	85.56	85.65	0.92	85.56	85.56
DCSwinB (no Conv-MLP)	88.56	87.02	86.15	0.93	86.35	86.68
DCSwinB (with Conv-MLP)	**90.96**	**90.56**	**89.65**	**0.94**	**85.56**	**87.95**

Bold values represent the best performance

Table 5 Parameters comparison DCSwinB vs. Swin-T

Model	Params (M)	FLOPs (G)	Inference Time (ms/img)	Top@acc(1%)
Swin-Tiny	87.94	85.56	85.65	0.92
DCSwinB (with Conv-MLP)	90.96	90.56	89.65	0.94

spatial enhancement and global dependency modeling enables DCSwinB to deliver both higher accuracy and greater reliability for lung nodule classification tasks.

To substantiate the claim of computational efficiency, we have conducted additional benchmarking and now report FLOPs, parameter count, and inference time comparisons between DCSwinB and the baseline Swin-Tiny model. as shown in Table 5. These results are based on experiments conducted on an NVIDIA RTX 4060 GPU with identical input size and batch settings.

DCSwinB reduces the number of parameters by approximately 19.8% compared to Swin-T. Our method achieves a 24.4% decrease in FLOPs, largely due to the use of depthwise convolutions and the dual-branch structure. On average, DCSwinB is 2.6 ms faster per image, representing a 16% improvement in inference speed.

Limitations and future work

The DCSwinB model introduces a dual-branch architecture, which enhances performance by capturing both local and global features. However, this results in increased computational complexity and memory usage. The model requires significant resources, which may limit its application in resource-constrained environments or on devices with limited computational power.The model's performance heavily relies on pretraining on datasets such as LUNA16 and LUNA16-K. While this pretraining helps the model generalize well, its performance might degrade when applied to datasets that are significantly different from those used in training.Further exploration of unsupervised

or self-supervised learning strategies could reduce dependency on pretrained models. Although the DCSwinB model performs well on the LUNA16 and LUNA16-K datasets, its generalization to other types of pulmonary nodules or CT imaging data remains uncertain. It may struggle to capture the full diversity of features present in different medical imaging datasets, particularly those with varied image quality, noise, or resolution.

Conclusion

This paper presents DCSwinB, a novel dual-branch Swin Transformer architecture designed to extract semantic contextual information from CT images with improved computational efficiency. By integrating Conv-MLP modules, DCSwinB strengthens connections between neighboring windows within the ViT branch, enhancing feature representation quality. Pretraining on LUNA16 and LUNA16-K datasets, followed by evaluation through ten-fold cross-validation, demonstrated the robustness and reliability of the proposed model. Experimental results show that DCSwinB achieves superior accuracy in differentiating benign from malignant pulmonary nodules compared to traditional feature-based and deep learning baselines. These findings suggest that DCSwinB offers a promising and efficient solution for enhancing early lung cancer detection, contributing to improved diagnostic outcomes in clinical practice. Future research will focus on extending DCSwinB to fully 3D volumetric data, incorporating multimodal information, and developing lightweight variants for broader clinical deployment.

Acknowledgements
We thank reviewers and editors for constructive and valuable advice for improving this article.

Authors' contributions
Writing-original draft preparation: Mohammad Khalid Faizi; review and editing: Juanjuan Zhao, Yan Qiang, Yangyang Wei, and Ying Qiao; data analysis: Rukhma Aftab, Zia Urrehman; supervision: Juanjuan Zhao.

Funding
This work is supported in part by the National Natural Science Foundation of China (NSFC) (Grant No. 62376183, U21 A20469, 62476190),
in part by the Central Government Guides Local Science and Technology development funding projects (Grant no. YDZJSX20220004),
in part by the special fund for Science and Technology Innovation Teams of Shanxi Province (Grant no. 202304051001009).

Data availability
The statement is the same as included in the main manuscript file. The Luna16 Dataset publicly available on: https://luna16.grand-challenge.org/.

Materials availability
Not applicable.

Code availability
After the acceptance of the paper, the code will be publicly open to access.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 October 2024 Accepted: 13 May 2025
Published online: 01 July 2025

References

1. Ferlay J, Colombet M, Soerjomataram I, Mathers CD, Parkin DM, Piñeros M, Znaor A, Bray F. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer. 2019;144(8):1941–1953.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71:209–49.
3. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransU-Net: Transformers Make Strong Encoders for Medical Image Segmentation. ArXiv. 2021;abs/2102.04306. https://arxiv.org/abs/2102.04306.
4. Lu Z, Wang Z, Huang D, Wu C, Liu X, Ouyang W, et al. FiT: Flexible Vision Transformer for Diffusion Model. ArXiv. 2024;abs/2402.12376. https://arxiv.org/abs/2402.12376.
5. Peng Y, Sonka M, Chen DZ. Group Vision Transformer. In: ACM Multimedia. 2024. https://api.semanticscholar.org/CorpusID:273645104. Accessed 15 May 2025.
6. Shi X, Hao Z, Yu Z. SpikingResformer: Bridging ResNet and Vision Transformer in Spiking Neural Networks. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. pp. 5610–9. Publisher: Institute of Electrical and Electronics Engineers (IEEE) Address: 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA
7. Oyovwe BE, Edje AE, Omede E, Ogeh C. An enhanced Convolutional Neural Network (CNN) model for the detection of lung cancer using X-Ray image. Sci Afr. 2024; 23(2):1–10. https://www.ajol.info/index.php/sa/article/view/270453.
8. Jafery NN, Sulaiman SN, Osman MK, Karim NKA, Soh ZHC. Enhancing Lung Cancer Classification: Leveraging Existing Convolutional Neural Networks within a 1D Framework. In: 2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE). 2024. pp. 52–7. Publisher: Institute of Electrical and Electronics Engineers (IEEE) Address: 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA
9. Gedam AN, Ajalkar DA, Rumale AS. Lung nodule detection using Eyrie Flock-based Deep Convolutional Neural Network. Intell Decis Technol. 2024;18:1651–73.
10. Liu B, Zhao X, Hu H, Lin Q, Huang J. Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. J Theory Pract Eng Sci. 2023;3(12):36–42. https://doi.org/10.53469/jtpes.2023.03(12).06.
11. Girshick RB. Fast R-CNN. 2015. https://api.semanticscholar.org/CorpusID:206770307. Accessed 15 May 2025.
12. Ren S, He K, Girshick RB, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2015;39:1137–49.
13. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal Speed and Accuracy of Object Detection. ArXiv. 2020;abs/2004.10934. https://arxiv.org/abs/2004.10934.
14. Calakli F, Taubin G. SSD: Smooth Signed Distance Surface Reconstruction. Comput Graph Forum. 2011;30(7):1993–2002.
15. Cheng X, Yu J. RetinaNet With Difference Channel Attention and Adaptively Spatial Feature Fusion for Steel Surface Defect Detection. IEEE Trans Instrum Meas. 2020;70:1–11.

16. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. CenterNet: Keypoint Triplets for Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019. pp. 6568–77. Publisher: Institute of Electrical and Electronics Engineers (IEEE) Address: 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA
17. Law H, Deng J. CornerNet: Detecting Objects as Paired Keypoints. Int J Comput Vis. 2018;128:642–56.
18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv. 2020;abs/2010.11929. https://arxiv.org/abs/2010.11929.
19. Rehman A, Mahmood T, Saba T. Robust kidney carcinoma prognosis and characterization using Swin-ViT and DeepLabV3+ with multi-model transfer learning. Appl Soft Comput. 2024;170:112518.
20. Liu S, Lin Y, Liu D, Wang P, Zhou B, Si F. Frequency-Enhanced Lightweight Vision Mamba Network for Medical Image Segmentation. IEEE Trans Instrum Meas. 2025;74:1–12.
21. Ma J, Li F, Wang B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. ArXiv. 2024;abs/2401.04722. https://arxiv.org/abs/2401.04722.
22. Wang Z, Zheng JQ, Zhang Y, Cui G, Li L. Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation. ArXiv. 2024;abs/2402.05079. https://arxiv.org/abs/2402.05079.
23. Wang Z, Ma C. Semi-Mamba-UNet: Pixel-Level Contrastive Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation. ArXiv. 2024;abs/2402.07245. https://arxiv.org/abs/2402.07245.
24. Wang Z, Ma C. Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better for Scribble-based Medical Image Segmentation. ArXiv. 2024;abs/2402.10887. https://arxiv.org/abs/2402.10887.
25. Ma C, Wang Z. Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual Mamba-based UNet for semi-supervised medical image segmentation. Knowl Based Syst. 2024;300:112203.
26. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, J'egou H. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. 2020. https://api.semanticscholar.org/CorpusID:229363322. Accessed 15 May 2025.
27. Yuan L, Chen Y, Wang T, Yu W, Shi Y, Tay FEH, et al. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021. pp. 538–47. Publisher: Institute of Electrical and Electronics Engineers (IEEE) Address: 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA
28. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in Transformer. In: Neural Information Processing Systems. 2021. https://api.semanticscholar.org/CorpusID:232076027. Accessed 15 May 2025.
29. Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J. Perceiver: General Perception with Iterative Attention. ArXiv. 2021;abs/2103.03206, 2021. https://arxiv.org/abs/2103.03206.
30. Zhang Z, Li X, Ma X, Sun Y. E-Transunet: Enhanced Transunet for Medical Image Segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). 2024. pp. 1–5. Publisher: Institute of Electrical and Electronics Engineers (IEEE) Address: 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA
31. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal. 2024;97:103280.
32. Eghbali N, Bagher-Ebadian H, Alhanai T, Ghassemi MM. GLoG-CSUnet: Enhancing Vision Transformers with Adaptable Radiomic Features for Medical Image Segmentation. 2025. https://api.semanticscholar.org/CorpusID:275336551. Accessed 15 May 2025.
33. Shah D, Khan MAU, Abrar M, Tahir M. Dual-View Deep Learning Model for Accurate Breast Cancer Detection in Mammograms. Int J Intell Syst. 2025;2025:1-17.
34. Shah D, Khan MAU, Abrar M, Tahir M. Optimizing Breast Cancer Detection With an Ensemble Deep Learning Approach. Int J Intell Syst. 2024;2024:1–17. https://doi.org/10.1155/2024/5564649.
35. Shah D, Khan MAU, Abrar M, Amin F, Alkhamees BF, Alsalman H. Enhancing the Quality and Authenticity of Synthetic Mammogram Images for Improved Breast Cancer Detection. IEEE Access. 2024;12:12189–98.
36. Shah D, Khan MAU, Abrar M. Reliable Breast Cancer Diagnosis with Deep Learning: DCGAN-Driven Mammogram Synthesis and Validity Assessment. Appl Comput Intell Soft Comput. 2024;2024:1122109:1-1122109:13.
37. Zhu W, Jin Y, Ma G, Chen G, Egger J, Zhang S, et al. Classification of lung cancer subtypes on CT images with synthetic pathological priors. Med Image Anal. 2023;95:103199.
38. Sua R, Shi R, Cui H, Xuan P, Fang C, Feng X, et al. TSEML: A task-specific embedding-based method for few-shot classification of cancer molecular subtypes. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2024. pp. 363–8. Publisher: Institute of Electrical and Electronics Engineers (IEEE) Address: 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA
39. Zhao H, Su Y, Lyu Z, Tian L, Xu P, Lin L, et al. Non-invasively Discriminating the Pathological Subtypes of Non-small Cell Lung Cancer with Pretreatment 18F-FDG PET/CT Using Deep Learning. Acad Radiol. 2023;30(11):1704–14. https://www.academicradiology.org/article/S1076-6332(23)00167-8/fulltext.
40. Wang N. Luna 16. IEEE Dataport. 2025. https://dx.doi.org/10.21227/0kjp-g187.
41. Mader KS. The Lung Image Database Consortium image collection (LIDC-IDRI). IEEE Dataport. 2021. https://dx.doi.org/10.21227/zce3-jp96.
42. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. ArXiv. 2019;abs/1912.01703, 2019. https://arxiv.org/abs/1912.01703.

Publisher's Note