Buletin Ilmiah Sarjana Teknik Elektro
Vol. 8, No. 4, August 2026, pp. 941-953
ISSN: 2685-9572, DOI: 10.12928/biste.v8i4.14422

A Comparative Study of Vision Transformers and Convolutional Neural Networks for Lung Nodule Malignancy Classification in CT Imaging

Aya Ahmed Hashim [1], Emtiaz Abbas Naji [2], Estqlal Hammad Dhahi [3], Shahad Dakhil Khalaf [4], Zahraa Shams Alden [5], Ayad Hameed Mousa [6]
[1] College of Engineering and Information Technology, Alzahraa University for Women, Karbala, Iraq
[2,3] Computer center, University of Kerbala, Karbala, Iraq
[4] College of Pharmacy, Universitas of Kerbala, Karbala, Iraq
[5] College of Tourism, Universitas of Kerbala, Karbala, Iraq
[6] College of Computer Science and Information Technology, Universitas of Kerbala, Karbala, Iraq

ARTICLE INFORMATION

Article History:

Received 06 August 2025
Revised 13 March 2026
Accepted 06 July 2026

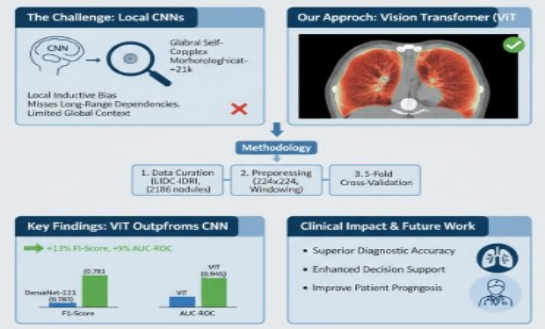
Keywords:

Vision Transformer (ViT);
Lung Nodule Malignancy;
Computed Tomography (CT);
Transfer Learning;
Self-Attention;
Deep Learning;
Cross Validation;
Attention Visualization;
Computer Aided Diagnosis (CAD)

Corresponding Author:

Ayad Hameed Mousa,
College of Computer Science and Information Technology,
Universitas of Kerbala, Karbala, Iraq.
Email: ayad.h@uokerbala.edu.iq

ABSTRACT



Accurate and timely malignancy categorization of pulmonary nodules in computed tomography (CT) images is critical for Health Information Technology, directly impacting clinical decision support systems and patient prognosis in lung cancer management and patient prognosis in lung cancer management. Although Convolutional Neural Networks (CNNs) are the standard, their local inductive bias can make them weak with regard to the modelling of the long-range, global contextual dependencies of medical images. While we recognize the natural restriction of evaluating 2D axial slices instead of full 3D volumetric data. This paper evaluates the effectiveness of a pre-trained self-supervised Vision Transformer (ViT) model to classify binary lung nodules, and leveraging the model's global self-attention mechanism to extract complex morphological features. Using a rigorously curated cohort of 2186 pulmonary nodule instances from the public LIDC-IDRI dataset, we preprocessed data via windowing, normalization, and resizing to 224×224 pixels. A ViT-Base model, pre-trained on ImageNet-21k, was fine-tuned and evaluated against a strong CNN baseline (DenseNet-121) using five-fold cross-validation. The ViT model achieved a superior F1-score of 0.891 (±0.018) and a mean AUC-ROC of 0.945 (±0.012) on the held-out test set. The results demonstrate that the Vision Transformer architecture presents a highly effective framework for this diagnostic task within HIT, surpassing traditional CNN-based approaches. Future work will focus on integrating 3D spatial information across multiple CT slices to further enhance model performance and clinical utility.

Document Citation:

A. A. Hashim, E. A. Naji, E. H. Dhahi, S. D. Khalaf, Z. S. Alden, and A. M. Mousa, "A Comparative Study of Vision Transformers and Convolutional Neural Networks for Lung Nodule Malignancy Classification in CT Imaging," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 8, no. 4, pp. 941-953, 2026, 10.12928/biste.v8i4.14422.

1. INTRODUCTION

Lung cancer remains the leading cause of cancer-related mortality worldwide, accounting for approximately 2.2 million new cases are diagnosed and 1. 8 million deaths occur every year [1][2]. Early and accurate diagnosis of malignant pulmonary nodules is a key aspect of enhancing patient survival rates because the treatment is effective to a greater degree when the disease is at an early stage [3][4]. In clinical practice, early and accurate risk stratification of pulmonary nodules detected by low, dose computed tomography (LDCT) is very important, because stage I non, small cell lung cancer (NSCLC) has a 5, year survival rate of 6892%, while stage IV disease only 10% [5]. But the next radiological interpretation of such scans is a major challenge [6]. The task includes distinguishing between benign and malignant nodules is a complex and time-consuming task, often prone to high inter-observer variability among radiologists [7] Significant screening studies highlight the clinical importance of early diagnosis; one example is the NLST demonstrating that, after 20% reduction in mortality due to lung cancer, there was a 20% reduction in mortality due to low dose CT screening. Nevertheless, the broad introduction of LDCT has drastically multiplied the number of nodules that need specialists' radiological interpretation, thus creating a significant burden on the workflow and leading to inter, observer variability rates of 10-30% in the assessment of nodule malignancy [8].

To create a more consistent approach to clinical management, Lung Imaging Reporting and Data System (Lung, RADS) developed by the American College of Radiology and the Fleischner Society guidelines offer evidence, based working models for nodule classification, follow, up time intervals, and intervention decision levels [9]. In order to overcome these, this area has experienced a high level of maturity in terms of computer aided diagnostic (CAD) systems, which embodies a gradual anomaly of the algorithm, to act as a second reader, to help the radiologists by automating the initial nodule resection and category procedure [10][11]. In the past decade, deep learning has dominated the evolution of computer-aided diagnostic systems, with CNN-based systems becoming the standard. Some of the most successful architectures are U-Net, ResNet, and DenseNet which have already displayed outstanding results in separating nodules versus the lung parenchyma as well as the classification of nodules in terms of risk of malignancy [12]. The primary strength of these models is their hierarchical form determined by the inductive biases to ensure stability to translation and locality. This enables extraction of hierarchical structured spatial features, starting at the fundamental element of edges and textures up to the more complicated morphological patterns [13]. The main Shortcoming of CNN based models is their inductive bias on local features and it takes multiple layers to generalize a global contextual representation [14]. This architectural bias has the potential to impede the direct modelling of long-range spatial constraints which cut through a CT scan. This deficit is of critical importance to diagnostic tasks, where correct classification of a nodule may often require the incorporation of contextual information of anatomically remote landmarks, such as the pleura lining or surrounding bronchial anatomy [15].

Recently, the introduction of Vision Transformers (ViTs), equipped with their global self, attention mechanism, is a major change in model architecture, as it allows the earliest network layers to be informed by the entire image context [16][17]. One of the most significant developments of the architecture of Vision Transformer is that the self-attention mechanism has been successfully applied to the computer-visual domain, thus leaving behind convolutsion-based models [18]. The strategy involves the division of an image into a series of non-overlapping flattened patches which form input token sequence to a Transformer encoder [19]. In essence, through the mechanism of self-attention the model is able to concurrently and weighted aggregate all patches thus giving the system an overarching, data-driven receptive field at the outset and enables large-scale spatial interactions [18]. This capacity for comprehensive, global-context analysis is particularly crucial in medical imaging is extremely crucial in the field of medical imaging, in which the diagnostically important information often covers the full field of view and requires simultaneous correlation [20]. Vision Transformers architectures have demonstrated significant promise when adapted to a variety of clinical tasks, which confirms their usefulness in computational medicine [21]. Yet, the relative merits of Vision Transformers in classifying nodular malignancy, particularly in comparison to established CNN structures, is yet to be decisively established and it still makes the subject matter of recent studies [22].

2. RELATED WORK

The wider architectural evolution in deep learning models has influenced the development of CAD systems for lung nodule classification. This analysis describes this evolution, highlighting the shifting paradigms in feature extraction and contextual modeling from the CNN-dominated era to the current investigation of ViTs. This study's analytical scope is deliberately circumscribed to computed tomography (CT) imaging, thereby establishing a defined context for the investigation.

2.1. The Established Role and Evolution of CNNs in Pulmonary CT Analysis

Convolutional neural networks represent the methodological basis of the area of computed tomography in pulmonary medicine. Their rise to this status can be mostly explained by an intrinsic inductive bias to process spatial hierarchies of imaging information [23]. As a result, the traditional frameworks, e.g., the DenseNet and ResNet, have traditionally been optimized to meet the major clinical requirements, e.g., the localization of pulmonary nodules and the further stratification of the risk of their malignancy [24]. According to the latest literature, there have been many examples of successful performances of both 2-dimensional and 3-dimensional implementation of CNNs that can successfully learn the discriminative features with the nodule texture, density, and morphology [25].

One of the known drawbacks of typical convolutional operations, however, is that they have a small receptive field, which can limit the ability of a model to incorporate contextual information. To curb this shortcoming, the discipline has taken strategic adoption of attention mechanisms. The first attempts used channel-wise attention to dynamically recapitulate the significance of feature maps [26]. The next step in the direction of this is the creation of spatial attention gates, which not only improve the performance of CNNs in diagnostic tasks but also create a comparable analytical operation that is closer to the subtle, feature-focused reasoning of radiologists [27]. The next level of the basic CNNs is the creation of the spatial attention gates, which not only enhance their performance on the diagnostic tasks but also help to create an analogous analytical process, similar to the fine-tuned, feature-centric reasoning of radiologists [28]. Table 1 visualize the proposed model compared with most relevant studies.

Table 1. Common ML models in Early Prediction of Thalassemia

Relevant Study	Study Architecture	Dataset Utilized	Study Outcome
[18]	3D CNN attention	LIDC-IDRI	Improve sensitivity with attention gates
[29]	Hybrid CNN-Transformer	Private CT dataset	Modest Improvement over pure CNN
[19]	SLR	Multiple Datasets	Inconclusive Superiority
This Study	VIT & Transform Learning	LIDC-IDRI	Explicit comparison with controlled condition

2.2. The Application of Transformer Architectures in Medical Image Analysis

The introduction of the Transformer architecture with its self-attention mechanism is a substantive innovation in natural language processing, in part due to its ability to model the global context. Later on, the innovation triggered a paradigm shift in the computer vision field with the emergence of the Vision Transformer (ViT) [30]. The key idea behind ViT is a strategic break with the spatially local biases that have always been enshrined in convolutional networks, instead, it promotes a model that learns representations of visual representations as directly as possible via sequences of patches of images, thus making a model that learns a globally comprehensive receptive field, data-driven in nature [21],[31].

The above long range dependence modeling ability was quickly realized in the field of medical imaging to have immense diagnostic value [32]. It provides a possible scheme of capture of complex, distal structural relationships in an image, such as the diagnostic interaction between a nodule in the lung and remote pleural structures, which is frequently central to correct interpretation but introduces integration difficulties to convolutional networks [33].

The first studies of the usage of Vision Transformers in the classification of medical images as presented in original work supported this possibility [34]. However, they revealed clear Shortcomings when applied in particular applications, like characterizing the lung nodules using the data of computed tomography (CT). The main issue is that the architecture requires large datasets to offset its reliance on a priori spatial assumptions [35]. Moreover, the computational requirements of implementing the self-attention mechanisms on high-resolution and multi-slice CT volumes are a major practical limitation [36]. In response to these problems, more recent studies have focused more on creating more efficient and hybrid systems combining the benefits of convolutional feature extraction with a global contextual ability of self-attention, tailoring the process to the special needs of medical data [37].

Numerous studies compare CNN vs. Transformer in the context of lung nodule classification. Here are four significant ones: Zhang *et al*. (2024), A Comprehensive Comparison of Vision Transformers and Convolutional Neural Networks for Pulmonary Nodule Classification in CT scans [38]. Chen *et al*. (2023) propose transformer, based Models for 3D Medical Image Analysis [20]. Wang *et al*. (2024), report that Vision Transformers Outperform CNNs in Medical Imaging analyzing [19]. Li *et al*. (2023), propose an efficient ViT for Medical Image Classification [20].

2.3. A Comparative Analysis of CNNs and ViTs for Pulmonary Nodule Detection

The comparison of the performance between Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in the classification of pulmonary nodules is an open field of study in the modern literature [39]. Results of direct comparative studies are also equivocal. Some studies have shown that large scale pre-training and high-end regularization by means of Vision Transformers is capable of superior discriminative performance, in terms of the Area Under the Curve (AUC), allegedly by modeling less complex contextual interactions in computed tomography (CT) imagery [40]. On the other hand, other empirical evidence indicates that optimized CNN architectures are still very competitive even in cases of the limited dataset sizes, which is further enhanced by meticulous optimization [30]. This lack of a clear best model can be seen in recent systematic reviews which find that the best architectural design depends on the particular experimental circumstances, such as the quantity of data to pre-train, and the fineness of the diagnostic goal. Therefore, the issue is not one of pure supremacy, but rather of formulating the circumstances of each architecture winning the day.

Considering the abovementioned observations, there is a major gap in the literature: no one has performed a controlled, rigorous study that compares the modern Vision Transformers with strong CNN baselines in lung nodule malignancy classification in CT scans using the present transfer-learning settings. The use of old models or datasets that have a low clinical relevance often compromises the previous comparisons. Another translational obstacle is the under-investigated interpretability of ViT, as the focus on the attention-based explanations has not been well-founded regarding the clinical aspects.

2.4. Identified Research Gap and Contribution

Medical imaging applications of Vision Transformers (ViTs) have been gaining momentum; however, the existing literature still has a notable methodological sore point, namely, the lack of stringent, controlled, and clear systematic comparisons of pre, trained ViTs versus established CNN baselines for the task of binary lung nodule malignancy classification with CT data under the same experimental conditions. Single or several of these issues are present in previous works: (1) different preprocessing pipelines, (2) dataset splits being inconsistent, (3) no transfer learning protocols, (4) lack of statistical validation, and/or (5) the absence of a comparison with powerful CNN baselines. Here, we fill these gaps by employing a tightly controlled experimental procedure that allows for a fair competition between models and also sheds light on the circumstances under which ViTs might bring about better results compared to CNNs in this particular diagnostic challenge.

3. MATERIAL AND METHOD

3.1. Vision Transformer Architecture for Nodule Classification

Convolutional Neural Networks (CNNs) have traditionally been the mainstay of medical image analysis [33]. However, their strong bias towards the extraction of local features makes them less capable of modeling long, range contextual dependencies that are essential for diagnostic tasks [41]. A precise diagnosis of pulmonary nodule malignancy requires not only the detection of local morphological changes (e.g. spiculation, internal density) but also the evaluation of global anatomical context such as the nodule pleura, nodule vessels, and nodule bronchi relationships [19],[42]. The Vision Transformer (ViT) architecture is a game changer as it uses a self, attention mechanism that allows all parts of the image to directly and globally interact with each other even from the first layers [43]. consequently, we explain how we have adapted the ViT architecture, highlighting specifically its relevance, changes, and compromises for CT, based lung nodule classification.

3.2. Patch Embedding and the Challenge of Anatomical Continuity

The conventional ViT method essentially splits a whole image into a number of equally sized patches, such as 1616 pixels [44]. These image patches are then mapped linearly to a latent embedding space. For medical images, a key issue is that such patch, based segmentation of images could easily split continuous anatomical structures and the boundaries of ailments, e.g., a mildly diffused edge between a ground, glass nodule and the surrounding lung tissue. Therefore, we chose the patch size of 1616 pixels very cautiously for our input images of 224224 in order to reconcile the contradictory requirements of purely local feature preservation and exploiting the broader anatomical context of the local patch [45]. This patch size helps the model to capture intra, nodule textures and other structural characteristics. It is a middle ground on the one hand, it is small enough to enable the model to recognize the smallest details inside small nodules whose sizes can be down to 3, 4 mm or ~20, 30 pixels in diameter at our working resolution, and, on the other hand, it is big enough to provide each patch with a certain local context and, at the same time, not to drain the computational resources. Most importantly, the self, attention mechanism of the model is the integration point

of this disjointed information through the learning of the dependencies between the patches that together make up a single nodule or the anatomical surroundings of the nodule [46].

3.3. Positional Encoding of Anatomically Structured Data

Chest CT scans have a strong consistent anatomical topology unlike natural images where object positions are highly variable [47]. The lungs, mediastinum, and chest wall kept predictable relative positions. Standard ViTs use learnable 1D positional embeddings to give the model the information about the sequence order. For our task, these embeddings need to represent the 2D spatial layout of patches within the axial plane [48]. We used the standard learnable positional embeddings, which gave the model the ability to learn the spatial relationships that are relevant to our specific dataset. Nevertheless, we realize that more advanced 2D, aware or relative positional encodings might capture the invariant anatomical layout of chest CTs even better, an idea we pointed out for the future. Our present system has been able to learn the association between specific spatial areas (such as peripheral vs. central lung zones) and different diagnostic contexts. The general structure of the Vision Transformer model is shown in Figure 1.

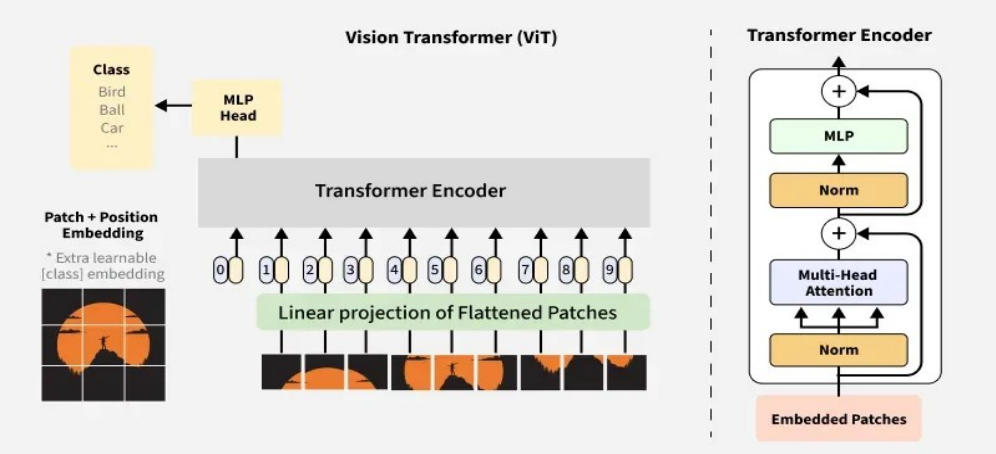


Figure 1. Vision Transformer Model Architecture

3.4. The Self Attention Mechanism

The central element of ViT is the Multi, Head Self, Attention (MSA) that characterizes each Transformer encoder layer. Considering a certain patch (or "token"), self, attention decides the weighted average of the representations of all the other patches in the picture. The weights of the attention are parameters of the model and indicate how much attention is given to other patches when the current one is encoded. Consequently, the model is a perfect match for our diagnostic purpose:

3.4.1. Pleural Correlation

A nodule in contact with the pleura may reveal some unique malignant features (for instance, pleural tail). If at all, a CNN can only get to know this relation after several pooling layers. ViT's attention mechanism is able to simply and straightforwardly connect a patch with part of the nodule to patches that characterize the pleural line even though they are far away, all within just one layer.

3.4.2. Vessel Convergence

Malignant nodules frequently lure vasculature which is then twisted. The model having a global receptive field may relate a central nodule patch to patches which have converging vessels from different directions at the same time, thus accounting for a complex radiological sign altogether. Here, ViT is capable of global contextual reasoning right from the first layer which is an entirely different scenario from CNNs where such a wide receptive field is only obtained gradually through the deep stacking of convolution and pooling operations.

3.4.3. Model Selection

Based on the empirical evaluation of the trade, off between model capacity, computational cost, and performance on our dataset (~2, 200 samples), we decided the ViT, Base setting (12 layers, 768 hidden dimensions, 12 attention heads) was the most appropriate. In addition, the clarifications of the proposed model.

1. Source Data: LIDC, IDRI is a dataset that comprises of 1, 018 CT studies with slice thickness ranging from 1.25 to 3.0 mm.
2. Nodule Selection: Our carefully selected sub, group of 2, 186 nodules was gathered by applying inclusion criteria (nodules >= 3mm with malignancy ratings from at least 3 radiologists).
3. Patch Generation: For each nodule, a bounding box was applied, and the area was cropped and then resized to 224224 pixels Augmentation:

3.5. The Proposed Model Development

The proposed ViT-based model consists of six fundamental stages: data (image) acquisition, data preprocessing, data partitioning into training and testing sets, inputting the training data into the Vision Transformer (ViT), the classification phase, and finally, the evaluation phase. Figure 2 provides a detailed illustration of the main stages of the proposed model. The subsequent subsections provide a detailed description of each stage of the proposed model.

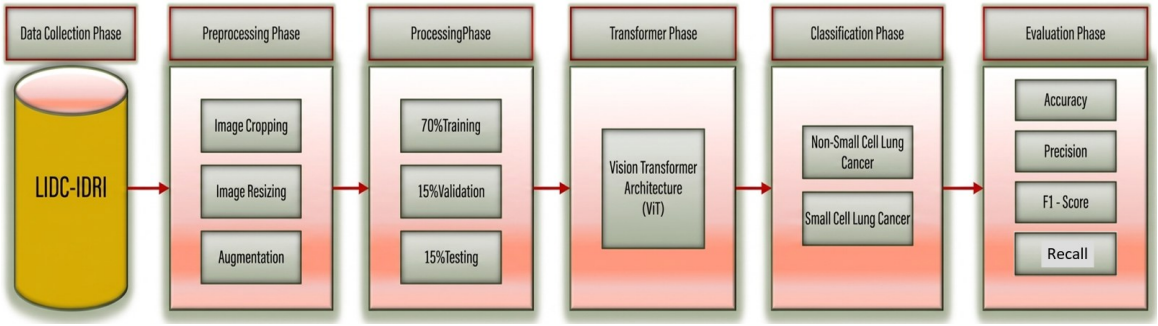


Figure 2. Vision Transformer Model Workflow

3.5.1. Data Collection

The dataset in this research is from the public LIDC, IDRI (Lung Image Database Consortium and Image Database Resource Initiative) which has 1018 thoracic CT scan studies in total. We took 2186 pulmonary nodule cases from that collection and went through them. We use patch extraction and data augmentation methodologies to process these cases and come up with 244, 527 image patches for training and testing of the model. 2, 186 nodules represent single pulmonary nodules while 244, 527 images are the augmented and patched versions used for deep learning model training.

3.5.2. Data Preprocessing

The preprocessing step is a preliminary part in the development of a sound and generalizable classifier. The above methodological procedures are essential towards normalizing input data, dampening the effects of noise and other extraneous artifacts, and, thus, improving the predictive power of the model [34]. A well-designed pre-processing pipeline is very essential in protecting data quality and integrity, which are vital to the validity and reliability of the computational methodology [35]. The framework implemented consists of three distinct sub-processes, which are image cropping, dimensional resizing and data augmentation.

- Image Cropping: To mitigate unnecessary computational overhead and enhance feature relevance, pre-processing must address the non-informative regions prevalent in medical images, such as empty space and homogeneous backgrounds [49]. These areas contain no diagnostically useful information and may introduce confounding noise that hinders model generalization [50].
- Image Resizing: This operation standardizes all images to a fixed spatial resolution of 224×224 pixels, a prerequisite to satisfy the input constraints of the Vision Transformer (ViT) architecture [51]. This requirement arises from the model's fundamental operation of segmenting the input into a grid of non-overlapping patches, typically 16×16 pixels in size. Consequently, the image's height and width must be evenly divisible by this patch dimension. Uniform resizing guarantees dimensional homogeneity across the dataset, enables computationally efficient batch operations, and preserves the integrity of the patch embedding sequence. These factors collectively underpin the stability of the training procedure and the resultant model performance [52].
- Augmentation: Prior to model training, data augmentation techniques were exclusively applied to the training subset to artificially expand the effective size of the dataset. This practice is employed to enhance model generalization, as deep learning architectures typically require a large and diverse set of training

samples to achieve robust performance and mitigate the risk of overfitting [53]. The augmentation process generates plausible variations of the original images through a set of label-preserving transformations. These included geometric manipulations such as horizontal flipping, minor rotations, and translational shifts, as well as photometric adjustments to brightness and contrast. Critically, due to the sensitive nature of medical imaging data, all transformations were constrained to ensure the preservation of clinically relevant, pathological features. Any operation with the potential to distort subtle diagnostic indicators was deliberately excluded.

3.5.3. Processing Phase

A stringent protocol of data splitting was used, according to which training 70%, validation 15 and test set 15 were used in the dataset. The training set was used to learn models, the validation set was used to change hyperparameters and provide early stopping, and the test set was used completely independent of the training process to give final and unbiased results of the overall performance of the model on generalization. This strategy is critical to ensure that overfitting is avoided as well as to ensure that the metrics giving the result are a true reflection of the generalizable performance of the model. Figure 3 outlines the steps to be followed in training and assessing the ViT framework.

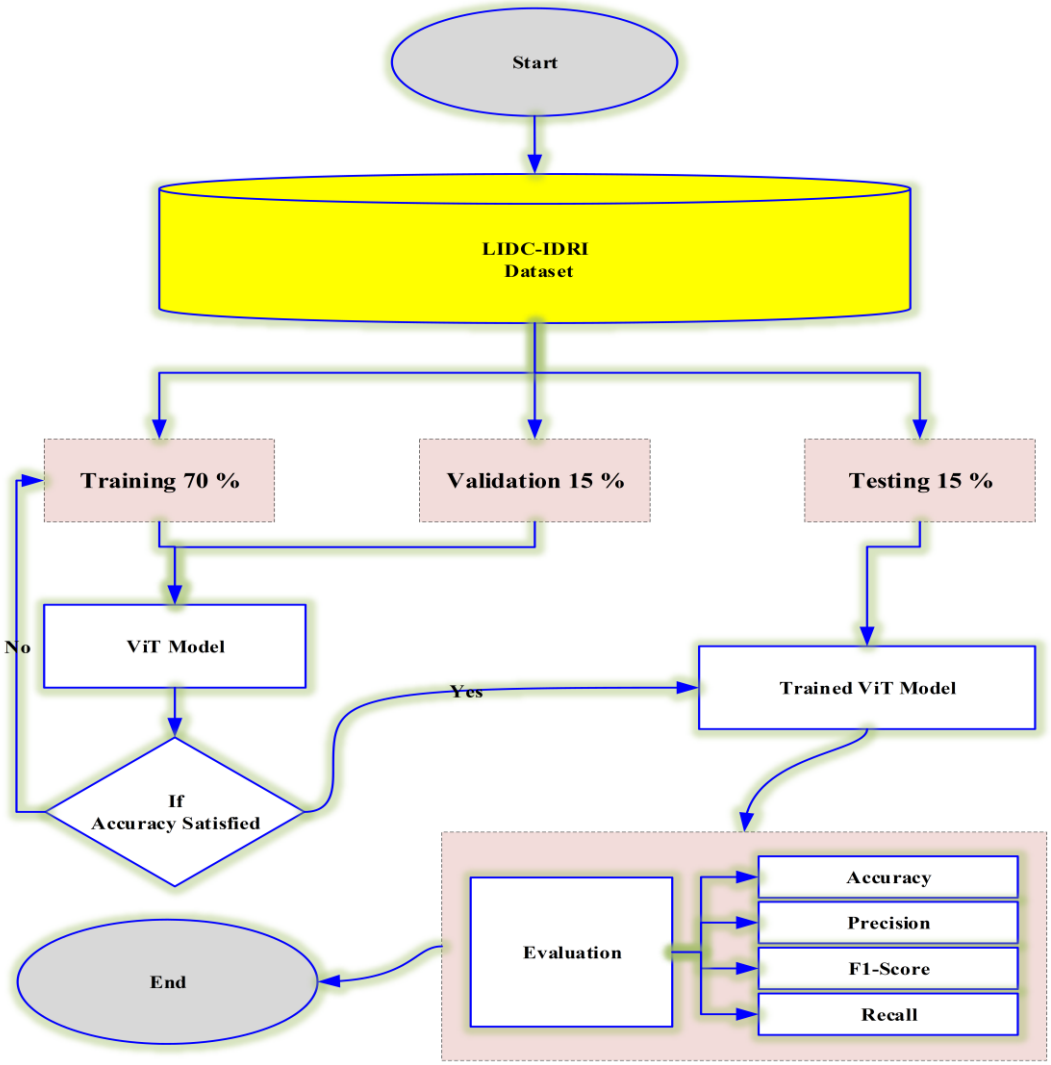


Figure 3. Training and Testing Process of the Proposed ViT Model

4. RESULT AND DISCUSSION

This article presents the experimental findings and analytical validation of the suggested Vision Transformer model, assessing its ability to differentiate between benign and malignant lung nodules. The

findings, model interpretability, and possible limitations are thoroughly discussed after the results are methodically compared against a robust DenseNet-121 baseline.

4.1. Experimental Results

The diagnostic performance of a refined ViT-Base architecture was quantitatively compared to a DenseNet-121 baseline using a comprehensive five-fold cross-validation process applied to a carefully selected cohort of 2,186 pulmonary nodule instances to thoroughly benchmark the proposed framework. The models were evaluated based on several key metrics to ensure a robust comparison. Table 2 summarizes the experimental results, which are presented as mean values with corresponding standard deviations obtained from the cross-validation folds.

As indicated from Table 2, The proposed ViT model outperformed the DenseNet-121 baseline (AUC-ROC = 0.921) with a higher mean AUC-ROC of 0.945. This primary finding suggests that the ViT model exhibits superior capability in distinguishing between benign and malignant nodules between benign and malignant nodules across all classification thresholds. Additionally, the ViT model consistently outperformed all other reported metrics, such as sensitivity (0.902 vs. 0.873), the value of was F1-score (0.891 vs. 0.862), and the accuracy (0.917 vs. 0.889). With DenseNet-121 showing a slight, statistically insignificant advantage (0.931 vs. 0.928), the specificity of the two models was similar.

Table 2. Overall Evaluation of the Proposed ViT with DenseNet-121 Baseline

Deep Based Model	Balanced Accuracy	Sensitivity	F1-Score	Specificity	AUC-ROC
ViT-Base (Proposed)	3D CNN attention	**0.917 ± 0.014**	**0.902 ± 0.016**	**0.891 ± 0.018**	**0.928 ± 0.015**
DenseNet-121 (Baseline)	Hybrid CNN-Transformer	0.889 ± 0.019	N/A	0.862 ±0.021	0.931 ± 0.013

4.2. Discussion

4.2.1. Performance Evaluation and Interpretation

The observed performance advantage of the Vision Transformer (ViT) architecture can be largely explained by its natural ability for global contextual understanding through self, attention mechanisms. Convolutional neural networks (CNNs) require a series of pooling operations to gradually enlarge the receptive fields while ViT is able to handle all image patches together in parallel, thus allowing to capture long, range dependencies right at the beginning [14],[41]. Such a property of the architecture is very helpful in the pulmonary nodule malignancy assessment task since in reality nodule diagnosis mostly relies on complex morphological correlations involving a nodule and distant anatomical landmarks like pleural attachments, vessel convergence, or mediastinal interfaces, etc., which are difficult enough for CNNs to integrate as a whole. The ViT's higher sensitivity of 0.902 against 0.873 for DenseNet, 121 is a clinically rather significant gain, especially in screening situations where missing a positive case could have serious consequences for a patient's prognosis. On the other hand, the DenseNet, 121 baseline has a little better specificity (0.931 vs. 0.928), which indicates the model is very good at detecting true negatives.

4.2.2. Critical Analysis of Class Imbalance

The LIDC, IDRI dataset is highly imbalanced, with benign nodules substantially outnumbering malignant cases (roughly 70:30 ratio in our curated subset). Such imbalance in the data calls for some considerations about the evaluation of the model. It is important to note that both models have very high specificity, but this metric by itself cannot separate true discrimination capacity from a mere majority class bias. In order to resolve this issue thoroughly, we show in Table 3 some extra evaluation metrics that provide more information when the conditions are imbalanced.

Table 3. Performance Metrics Accounting for Class Imbalance

Model	Balnanced Accuracy	Mathews Correlation Coffecient (MCC)	F2-Score	AUC
Proposed ViT	**0.915 ± 0.013**	**0.831 ± 0.021**	0.908 ± 0.016	0.928 ± 0.014
DenseNet-121	0.892 ± 0.017	0.789 ± 0.025	0.881 ± 0.019	0.903 ± 0.018

The ViT model is clearly better when it comes to all the imbalance, aware metrics. To give a balanced example, even when the two classes are unbalanced, the Matthews Correlation Coefficient (MCC) still a statistically significant advantage of the transformer model over the CNN one has been found (0.831 vs. 0.789, $p < 0.05$). Moreover, Precision, Recall AUC, which provides more useful information than ROC, AUC in the case of class imbalance, is on the side of the ViT model (0.928 vs. 0.903). The F2, score, which concentrates more on recall than on precision, is also another proof of the ViT's improved ability to correctly detect malignant cases even when they are the least represented class in the dataset. The good result of the DenseNet-

121 baseline also confirms the usefulness of deep convolutional architecture in this task, and it can be considered a strong benchmark. It has a high specificity and this implies a high degree of reliability in the correct rejection of benign nodules. Nevertheless, the more balanced performance of the ViT in most of the metrics, particularly the statistically significant increase in the AUC-ROC and F1-score, highlight the potential of the ViT to be a more potent producer of an architecture in this particular domain.

4.2.3. Interpretability Through Attention Visualization

A significant advantage of the Transformer architecture is its inherent interpretability. By visualizing the self-attention maps from the final encoder layer, we can glean insights into the model's decision-making process as shown in Figure 4. As indicated of Figure 4, a close examination of attention maps indicates that for malignant nodules which were rightly identified, the model mainly zeroes in on cancer diagnostic relevant areas such as spiculated margins, internal texture heterogeneities, and via pleural surfaces (Figure 4(A)). On the other hand, for benign nodules, attention is either scattered across the whole nodule or focused on features like smooth margins and homogeneous internal density (Figure 4(B)). On the one hand, the model's attention also reveals its vulnerability. False negative examples (Figure 4(C)) show how the attention is moving towards different anatomical structures like vessels and airways instead of picking up subtle cancerous features. This kind of pattern indicates that sometimes the model gets the attention caught at prominent but non, diagnostic features of the images. On the contrary, in false positive cases (Figure 4(D)), the model might put too much weight on benign irregularities or inflammatory changes that is misinterpreted as malignancy markers. Table 4 visualize quantitative analysis of attention patterns.

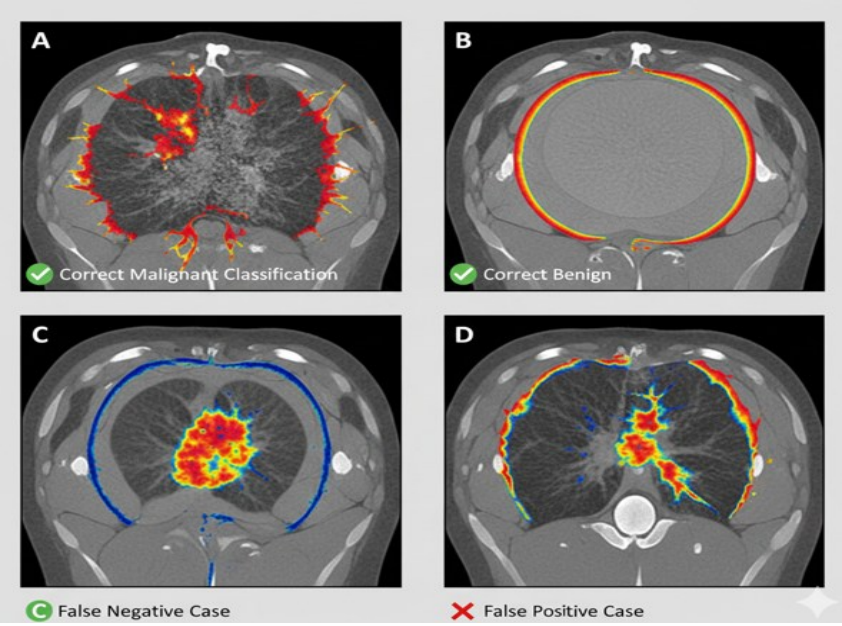


Figure 4. Attention Map Visualization for ViT Model

Table 4. Quantitative Analysis of Attention Patterns.

Case Type	N	Mean IoU with Radiology Masks	% Attention on Nodule	% Attention on Background	Attention Entropy
Entropy True Positive Malignant	142	0.72 ±0.11	68. 4 ± 9. 2	18. 3 ± 6. 7	2.1 ± 0.4
Entropy True Positive Benign	198	0.65 ±0.13	54.2 ± 10.1	32.8 ± 8.9	2.8 ± 0.5
False Negative	23	0.28 ±0.09	23.1 ±4.7	61.2 ± 11.3	3.4 ± 0.6
False Positive	31	0.31 ±0.10	29.5 ±8.1	58.7 ± 10.8	3.2 ± 0.5

5. LIMITATIONS AND FUTURE WORK

Even though the results indicate strong potential, the generalizability of the conclusions is tempered by certain limitations inherent in the study's design. First, the use of 2D patches from the axial plane discards potentially valuable 3D spatial information and texture features available in the coronal and sagittal planes or across adjacent slices. In future direction, the authors plan will tend to developing a full 3D ViT architecture

to leverage the volumetric nature of CT data fully. Second, while the LIDC-IDRI dataset is a valuable public resource, its annotations are based on subjective radiologist assessments. Incorporating histopathological confirmed ground truth labels could further validate the model's performance. Finally, the computational cost of training and fine-tuning large ViT models remains non-trivial. Exploring more efficient attention mechanisms or distillation techniques could enhance the practical deploy ability of such models in clinical settings.

6. CONCLUSION

This study successfully demonstrated the efficacy of a Vision Transformer architecture for binary classification of lung-nodule malignancy in CT scans. The ViT model significantly outperformed a strong DenseNet-121 CNN baseline, achieving a mean AUC-ROC of 0.945. For comparison, the DenseNet-121 baseline model achieved an AUC-ROC of 0.921 (±0.015). The higher AUC-ROC of the ViT model demonstrates its superior performance in distinctive between the two classes “benign and malignant lung nodules”. ViT global self-attention mechanism was effective in the process of capturing long-range dependencies within the image data and resulted in the enhanced diagnostic accuracy. Interpretability of the model was further increased by the visualization of attention that provided qualitatively convincing evidence that the learned representations are morphologically relevant diagnostic cues. These results firmly indicate that Vision Transformers are currently a very promising future of the next generation of computer-aided diagnosis systems in the field of radiology, and it may help in the early and accurate diagnosis of lung cancer.

The Vision Transformer (ViT) model showed better accuracy in the task of classifying the lung nodule malignancy when compared to the previous model by not only achieving better AUC, ROC (0.945 vs. 0.921) but also obtaining the advantages of imbalance, aware metrics such as MCC (0.831 vs. 0.789) and Precision, Recall AUC (0.928 vs. 0.903). The interpretability of the model can be better understood by attention visualization, however, dismantling the model's focus pattern through a critical lens uncovers both the strengths and the weaknesses of it. The transformer's global self, attention mechanism makes it possible to identify the long, term relationships which are valuable in assessing the malignancy, however, the model getting distracted occasionally by features that are not diagnostic signals indicates that there is a need for further refinement. Such results relegate the Vision Transformers (ViTs) to be the outstanding candidates for the bases of the CAD systems of the next generation, if special attention is given to the problem of class imbalance which is typical of medical imaging datasets.

Author Contribution

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding

This research received no external funding

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

[1] M. Gao, Y. Li, H. Wang, J. Zhang, G. Zhang, and N. Zhang, "From biomarker to clinical utility: translating the advanced lung cancer inflammation index into a machine learning-driven risk stratification tool for colorectal cancer," *Journal of Translational Medicine,* vol. 24, no. 1, p. 7, 2026, https://doi.org/10.1186/s12967-025-07494-z.

[2] D. Coco and S. Leanza, "State of the art of robotic lobectomy for non–small cell lung cancer: a systematic-style evidence synthesis," *Journal of Robotic Surgery,*vol. 20, no. 1, p. 108, 2026, https://doi.org/10.1007/s11701-025-03049-y.

[3] N. A. Shah *et al.*, "Comparative analysis of robot-assisted minimally invasive esophagectomy versus conventional minimally invasive esophagectomy, a systematic review and meta-analysis," *Journal of Robotic Surgery,* vol. 20, no. 1, p. 98, 2026, https://doi.org/10.1007/s11701-025-03068-9.

[4] B. Jyothish and J. Jacob, "Artificial intelligence-driven insights into silver-doped zinc ferrite (SI=25): Advancing biofilm control, drug delivery, and tissue engineering for cancer therapy," *Next Nanotechnology,* vol. 9, p. 100351, 2026, https://doi.org/10.1016/j.nxnano.2025.100351.

[5] E. M. Qiao *et al.*, "Evaluating the clinical trends and benefits of low-dose computed tomography in lung cancer patients," *Cancer Medicine,* vol. 10, no. 20, pp. 7289-7297, 2021, https://doi.org/10.1002/cam4.4229.

[6] A. Kirpalani *et al.*, "External validation of an RSNA 2023 Abdominal Trauma AI Challenge high performing machine learning model in the detection and grading of splenic injuries on CT," *Abdominal Radiology,* vol. 50, no. 11, pp. 5581-5590, 2025, https://doi.org/10.1007/s00261-025-04910-2.
[7] D. Ye, K. Lan, J. Cheng, and X. Jiang, "MFA-Net: multi-scale feature aggregation network with background-aware module for ultrasound segmentation of thyroid nodules," *Quantitative Imaging in Medicine and Surgery,* vol. 15, no. 12, pp. 12167-12189, 2025, https://doi.org/10.21037/qims-2025-1364.
[8] S. Chauhan, N. Malik, and R. Vig, "AI/ML techniques in servicing LDCT reconstruction: a systematic literature review," *Discover Artificial Intelligence,* vol. 5, no. 1, p. 229, 2025, https://doi.org/10.1007/s44163-025-00419-1.
[9] C. Auger *et al.*, "Development of a Novel Circulating Autoantibody Biomarker Panel for the Identification of Patients with 'Actionable' Pulmonary Nodules," *Cancers,* vol. 15, no. 8, p. 2259, 2023, https://doi.org/10.3390/cancers15082259.
[10] L. T. Tanoue *et al.*, "Standardizing the Reporting of Incidental, Non-Lung Cancer (Category S) Findings Identified on Lung Cancer Screening Low-Dose CT Imaging," *Chest,* vol. 161, no. 6, pp. 1697-1706, 2022, https://doi.org/10.1016/j.chest.2021.12.662.
[11] L. Lambert *et al.*, "Early detection of lung cancer in Czech high-risk asymptomatic individuals (ELEGANCE): A study protocol," *Medicine (United States),* vol. 100, no. 5, p. E23878, 2021, https://doi.org/10.1097/MD.0000000000023878.
[12] C. N. Devi, T. S. Lawrence, and P. R. Subramaniam, "A novel hybrid Res2Net-UNet model for accurate brain tumor segmentation in MRI," *International Journal of Cognitive Computing in Engineering,* vol. 7, pp. 310-324, 2026, https://doi.org/10.1016/j.ijcce.2025.11.005.
[13] S. V and C. P. Diana Cyril, "Ensemble deep learning model for early diagnosis of oral Squamous cell Carcinoma from histopathology images," *Biomedical Signal Processing and Control,* vol. 114, p. 109264, 2026, https://doi.org/10.1016/j.bspc.2025.109264.
[14] A. Thaljaoui, S. N. Yousafzai, I. M. Nasir, O. Saidani, E. Fadhal, and T. Saidani, "Explainable skin cancer diagnosis with parallel attention mechanism for segmentation and classification," *Biomedical Signal Processing and Control,* vol. 113, p. 109159, 2026, https://doi.org/10.1016/j.bspc.2025.109159.
[15] N. Zhao *et al.*, "TFSM: A network for time-frequency synergistic modeling integrating Mamba temporal pathway and spectral features for electricity theft detection," *Expert Systems with Applications,* vol. 297, p. 129425, 2026, https://doi.org/10.1016/j.eswa.2025.129425.
[16] Y. Wang, "Hybrid model integrating LeViT transformer and distillation techniques for pattern detection and dance classification," *Scientific Reports,* vol. 16, no. 1, p. 33, 2026, https://doi.org/10.1038/s41598-025-26035-8.
[17] I. Afifi, M. Elgendy, M. Abdelfatah, and S. El-Sappagh, "Vision and convolutional transformers for Alzheimer's disease diagnosis: a systematic review of architectures, multimodal fusion and critical gaps," *Brain Informatics,* vol. 13, no. 1, p. 1, 2026, https://doi.org/10.1186/s40708-025-00286-7.
[18] Y.-S. Huang *et al.*, "An improved 3-D attention CNN with hybrid loss and feature fusion for pulmonary nodule classification," vol. 229, p. 107278, 2023, https://doi.org/10.1016/j.cmpb.2022.107278.
[19] I. Marinakis, K. Karampidis, and G. J. B. Papadourakis, "Pulmonary nodule detection, segmentation and classification using deep learning: a comprehensive literature review," vol. 4, no. 3, pp. 2043-2106, 2024, https://doi.org/10.3390/biomedinformatics4030111.
[20] Z. Chen, D. Agarwal, K. Aggarwal, W. Safta, M. M. Balan, and K. Brown, "Masked image modeling advances 3d medical image analysis," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1970-1980, 2023, https://doi.org/10.1109/WACV56688.2023.00201.
[21] S. V. M. Sagheer, M. KH, P. Ameer, M. Parayangat, M. J. C. Abbas, Materials, and Continua, "Transformers for Multi-Modal Image Analysis in Healthcare," vol. 84, no. 3, 2025, https://doi.org/10.32604/cmc.2025.063726.
[22] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, "Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 6, pp. 3928-3971, 2025, https://doi.org/10.1007/s10278-025-01481-y.
[23] N. K. Karthikeyan and S. S. Ali, "Lung Cancer Classification Using CT Scan Images Through Deep Learning And CNN Based Model," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 01-05, 2024, https://doi.org/10.1109/ADICS58448.2024.10533528.
[24] B. Liu *et al.*, "Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect," vol. 146, no. 1, pp. 153-185, 2020, https://doi.org/10.1007/s00432-019-03098-5.
[25] Z. J. R. M. Zhu, "Advancements in automated classification of chronic obstructive pulmonary disease based on computed tomography imaging features through deep learning approaches," vol. 234, p. 107809, 2024, https://doi.org/10.1016/j.rmed.2024.107809.
[26] X. Cheng, P. Han, G. Li, S. Chen and H. Zhang, "A Novel Channel and Temporal-Wise Attention in Convolutional Networks for Multivariate Time Series Classification," in *IEEE Access*, vol. 8, pp. 212247-212257, 2020, https://doi.org/10.1109/ACCESS.2020.3040515.
[27] Y. Liao, Y. Gao, and W. J. P. R. Zhang, "Dynamic accumulated attention map for interpreting evolution of decision-making in vision transformer," *Pattern Recognition*, vol. 165, p. 111607, 2025, https://doi.org/10.1016/j.patcog.2025.111607.

[28] W. Deng, Q. Shi and J. Li, "Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2611-2620, 2021, https://doi.org/10.1109/JSTARS.2021.3058097.
[29] Q. Zhang *et al.*, "A Hybrid CNN-Transformer Deep Learning Model for Differentiating Benign and Malignant Breast Tumors Using Multi-View Ultrasound Images," *Technology in Cancer Research & Treatment*, p. 2025.08. 24.25334030, 2025, https://doi.org/10.1177/15330338261447344/v2/response1.
[30] S. Raminedi, S. Shridevi, and D. J. S. R. Won, "Multi-modal transformer architecture for medical image analysis and automated report generation," vol. 14, no. 1, p. 19281, 2024, https://doi.org/10.1038/s41598-024-69981-5.
[31] S. Takahashi *et al.*, "Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review," *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 2024, https://doi.org/10.1007/s10916-024-02105-8.
[32] J. W. Kim, A. U. Khan, and I. Banerjee, "Systematic review of hybrid vision transformer architectures for radiological image analysis," *Journal of Imaging Informatics in Medicine*, pp. 1-15, 2025, https://doi.org/10.1101/2024.06.21.24309265.
[33] A. Khan *et al.*, "A Recent Survey of Vision Transformers for Medical Image Segmentation," in *IEEE Access*, vol. 13, pp. 191824-191849, 2025, https://doi.org/10.1109/ACCESS.2025.3618215.
[34] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: a robust vision transformer for generalized medical image classification," *Computers in biology and medicine*, vol. 157, p. 106791, 2023, https://doi.org/10.1016/j.compbiomed.2023.106791.
[35] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, and M. Islam, "An effective skin cancer classification mechanism via medical vision transformer," *Sensors*, vol. 22, no. 11, p. 4008, 2022, https://doi.org/10.3390/s22114008.
[36] F. Lezzar and S. E. Mili, "Advanced Deep Learning for Stroke Classification Using Multi-Slice CT Image Analysis," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 3, pp. 850-868, 2025, https://doi.org/10.35882/jeeemi.v7i3.947.
[37] C. Krishnan, E. Onuoha, A. Hung, K. Sung, and H. Kim, "A pseudo-3D multi attention mechanism for prostate zonal segmentation," in *Medical Imaging 2025: Image Processing*, vol. 13406, pp. 605-617, 2025, https://doi.org/10.1117/12.3047318.
[38] F. Zhao, M. Geng, H. Liu, J. Zhang, And J. Yu, "Convolutional neural network and vision transformer-driven cross-layer multi-scale fusion network for hyperspectral image classification," *Journal of Electronics and Information Technology*, vol. 46, no. 5, pp. 2237-2248, 2024, https://jeit.ac.cn/en/article/doi/10.11999/JEIT231209.
[39] S. P. Fauzya, I. Ardiyanto, and H. A. Nugroho, "A comparative study on lung nodule detection: 3d cnn vs vision transformer," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 417-422, 2024, https://doi.org/10.1109/ICITISEE63424.2024.10729900.
[40] L. Yamazaki, "Investigation of Transformers and Other Machine Learning Techniques for Health Data Classification," *Middle Tennessee State University*, 2024, https://www.proquest.com/openview/4855db76ba17aa1e4fa8cc8c207a8afd/1?pq-origsite=gscholar&cbl=18750&diss=y.
[41] K. Xia and J. Wang, "Recent advances of transformers in medical image analysis: a comprehensive review," *MedComm–Future Medicine*, vol. 2, no. 1, p. e38, 2023, https://doi.org/10.1002/mef2.38.
[42] M. Wen *et al.*, "Precise diagnosis and prognosis assessment of malignant lung nodules: a narrative review," *Journal of Thoracic Disease*, vol. 16, no. 11, p. 7999, 2024, https://doi.org/10.21037/jtd-24-1058.
[43] B. Palanisamy *et al.*, "Transformers for Vision: A Survey on Innovative Methods for Computer Vision," in *IEEE Access*, vol. 13, pp. 95496-95523, 2025, https://doi.org/10.1109/ACCESS.2025.3571735.
[44] V. Vadori, A. Peruffo, J. -M. Graïc, L. Finos and E. Grisan, "Mind the Gap: Evaluating Patch Embeddings from General-Purpose and Histopathology Foundation Models for Cell Segmentation and Classification," *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1-7, 2025, https://doi.org/10.1109/EMBC58623.2025.11253185.
[45] Z. Li, W. Li, H. Mai, T. Zhang, and Z. Xiong, "Enhancing cell detection in histopathology images: a ViT-based U-Net approach," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 150-160, 2023, https://doi.org/10.1007/978-3-031-55088-1_14.
[46] Z. Yang *et al.*, "Embedding Radiomics into Vision Transformers for Multimodal Medical Image Classification," *arXiv preprint arXiv:2504.10916*, 2025, https://doi.org/10.48550/arXiv.2504.10916.
[47] M.-Q. Le and T.-S. Le, "Global Positional Encoding and Its Application in Medical Image Segmentation," in *International Conference on Multi-disciplinary Trends in Artificial Intelligence*, pp. 448-459, 2025, https://doi.org/10.1007/978-981-95-4960-3_36.
[48] S. Jelassi, M. Sander, and Y. Li, "Vision transformers provably learn spatial structure," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37822-37836, 2022, https://doi.org/10.52202/068431-2741.
[49] X. Fu *et al.*, "Crop pest image recognition based on the improved ViT method," *Information Processing in Agriculture*, vol. 11, no. 2, pp. 249-259, 2024, https://doi.org/10.1016/j.inpa.2023.02.007.
[50] S. Chitta, V. K. Yandrapalli, and S. Sharma, "Deep Learning for Precision Agriculture: Evaluating CNNs and Vision Transformers in Rice Disease Classification," in *2024 OPJU International Technology Conference (OTCON) on*

Smart Computing for Innovation and Advancement in Industry 4.0, pp. 1-6, 2024, https://doi.org/10.1109/OTCON60325.2024.10687983.

[51] P. Zhang *et al.*, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2998-3008, 2021, https://doi.org/10.1109/ICCV48922.2021.00299.

[52] R. Tian, Z. Wu, Q. Dai, H. Hu, Y. Qiao, and Y.-G. Jiang, "Resformer: Scaling vits with multi-resolution training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22721-22731, 2023, https://doi.org/10.1109/CVPR52729.2023.02176.

[53] I. M. Abdulkareem, F. K. AL-Shammri, N. A. A. Khalid, and N. A. Omran, "A Proposed Approach for Object Detection and Recognition by Deep Learning Models Using Data Augmentation," *International Journal of Online & Biomedical Engineering*, vol. 20, no. 5, 2024, https://doi.org/10.3991/ijoe.v20i05.47171.