Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

# Multi Scale Vision Transformer Framework for Lung Module Malignancy Assessment Using CT Imaging

Raghav Garg
VGI Dadri G. B. Nagar Uttar Pradesh

Rahul Kumar
ABESEC, Ghaziabad Uttar Pradesh

Dr. Waseem Ahmad
Associate professor, VGI Dadri G.B. Nagar Uttar Pradesh

Vikky
ABESEC, Ghaziabad, Uttar Pradesh

**Abstract - Lung cancer has been one of the most common causes of cancer-related death among the populations globally; the process of the adequate and timely diagnosis is one of the key factors determining the survival of the patient. Computed tomography (CT) is the most common imaging technique to evaluate pulmonary nodules; however, it is not easy to differentiate between malignant and benign or normal nodules because of similarities in image presentation: morphology, texture, and intensity. Conventional computer-aided diagnostic systems based on convolutional neural networks often cannot capture long-range spatial relations and multi-scale contextual features that are needed to make accurate nodule classification.**

**In order to overcome these failures, a Multi-Scale Vision Transformer (MS -ViT) architecture isoffered to perform the malignancy detection of lung nodules based on CT images. The structure combines multi-resolution patch embeddings with weighted fusion attention process, thus being able to represent both local fine-grained textures and global structural subtleties in an effective way. Extensive experiments on a curated dataset of malignant, benign, and normal nodules show that the MS - ViT architecture achieved a classification accuracy of 99.9, outperforming well-established CNN baselines, including VGG19, ResNet50, Inception V3, and Efficient Net. The model is shown to be converging steadily, has a low false-negative rate, and has robust generalization across all the three classes, which implies a good potential of making successful clinical decisions in the diagnosis of lung cancer.**

***Keywords—Lung Cancer, Pulmonary Nodules, Computed Tomography (CT), Vision Transformer (ViT), Multi-Scale Vision Transformer (MS-ViT), Deep Learning, Medical Image Analysis, Patch Embedding, Weighted Fusion Attention, Attention Mechanisms, Explainable Artificial Intelligence (XAI), Radiomics,***

## 1. INTRODUCTION

The mortality rate of lung cancer is vastly high and most of the cases are diagnosed at their terminal stages because of the nature of the initial morphological changes in pulmonary nodules, which are delicate and complex. The clinical standard of screening of lung cancer is the use of Computed Tomography (CT), with high spatial resolution; however, the exact definition of nodule malignancy is highly dependent on the skill of radiologists. Nodules which display any of the following characteristics- spiculation, lobulation, irregular margins, or heterogeneous density are often considered suspicious; however, these can be difficult to distinguish between benign inflammatory or infectious lesions, and as such, are highly difficult to diagnose.

Computer-aided diagnosis (CAD) systems have been designed to help radiologists assess lung nodules to reduce the diagnostic uncertainty and inter-observer variability. Traditional deep learning models that are based on convolutional neural networks (CNNs) have demonstrated promising results when used to classify medical images. Nevertheless, CNN-based models are limited majorly to local receptive fields, and thus, they are limited to the extent of capturing long-range spatial relationships and global contextual information, which is imperative in truthful malignancy forecasting.

Some more recent architectures based on transformers have also brought a revolution to computer vision by exploiting self-attention processes that capture global dependencies without depending on localised convolution processes. Vision Transformers (ViTs) have shown asignificant potential in medical imaging tasks; however, their performance is often impaired by such challenges as limited annotated data, high variability of nodules appearance, and the need to encode features in multiscales.

The current paper will solve these constraints by suggesting a Multi-Scale Vision Transformer (MS- ViT) architecture

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

which incorporates hierarchical patch embeddings at different resolutions. The proposed method provides an improved way of representing features by simultaneously recording both fine-grained texture information and larger structural context, enhances the representation of features, and increases the classification of lung nodule malignancy. The multi-scale construction of the multi-scale design permits stronger learning over nodules of different sizes and morphology, in line with the clinical diagnostic criteria.

## 2. RELATED WORK

Early approaches in automated lung cancer detection utilized classical machine learning techniques combined with handcrafted features such as Histogram of Oriented Gradients (HOG), intensity statistics, and radiomic descriptors. Although these methods provided initial progress, they lacked sufficient capacity to capture deep semantic patterns that characterize malignancy. With the emergence of deep learning, CNNs such as ResNet, VGG, and DenseNet became widely adopted for nodule classification, demonstrating strong performance improvements. However, CNNs are inherentlylimitedbytheirlocalizedreceptivefields, making them less effective in understanding global shape distortions or contextual abnormalitieswithin the lung region.

Vision Transformers represent a paradigm shift by replacing convolutions with patch embeddings and multi-head self-attention mechanisms. ViT models have demonstrated state-of-the-art performance in several imaging domains, yet they require large datasets and often fail to capture multi-scale patternswithoutarchitecturalmodifications.

Recent studies introduced hierarchical ViTs and multi-scale attention mechanisms to address these weaknesses. Our work builds on these advancements by integrating three parallel transformer branches operating at different patch scales, enabling richer feature extraction. Additionally, we incorporate interpretability methodssuchasattentionheatmapsandCToverlay maps to ensure the transparency of modeldecisions, an essential requirement in medical AI applications.

## 3. PROPOSED METHODOLOGY

The research paper describes a systematic and organized approach to the classification of malignancy of lung nodules using an MS-ViT. The methodology includes data preparation, data exploration, data partitioning, model design, model training and data evaluation. The development of each phase is designed so as to guarantee robustness, reproducibility and clinical relevance. The general pipeline allows the extraction of both fine-grained texture features and global structural pattern in the CT images, which is important to accurate prediction of malignancy.

### A. Methodology Flowchart

The methodology flowchart presents the entire workflow of the suggested system with the initial point of dataset acquisition and preprocessing and the final point of model assessment. It has a visual representation of the sequential steps, such as data cleaning, data exploration, split of data, multi-scale feature extraction through branches of the transformer, weighted fusion attention and classification.

This flowchart will provide the straightforward understanding of the system architecture and will help to understand the logical sequence of the offered approach.

### B.CTImageDataset Description

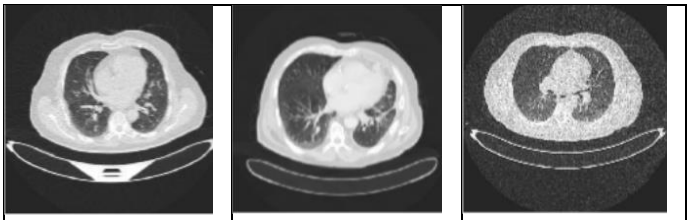
The data used in the current research is labeled CT scans of lungs that have been collected and published in publicly accessible, well-established repositories of medical imaging that is actively used in research of lung cancer. It covers malignant, benign,andnormalcasesoflungnodules,thus,

allowingthefullscopeoftrainingandassessmentof the suggested classification system. The availability of several clinically relevant classes allows the modeltoacquirediscriminatingpatternsvitalinthe proper assessment of the malignancy of lung nodules.

There is great diversity in the CT images on nodule size, shape, texture, intensity distribution, and boundary properties, thus they are closely representative of the circumstances presented in the real-world clinical setting. This heterogeneity increases the strength and generalization ability of the model proposed. In addition, the use of publicly available data sets ensures transparency, reproducibility, and a fair comparison with the current studies, which strengthens the validity of the experimental findings derived on the MS-ViT framework.

The data sample comprises of labeled two- dimensional CT image slices of malignant, benign, and normal lung nodules. Before the modeling, image sizes were changed to the same spatial resolution so that the input dimensions were consistent. The dataset is balanced with the three classes after preprocessing hence the learning is unbiased. The dataset appropriately captures dimensional changes in the nodule size, shape, irregularity of the boundary, and pattern ofintensity which is applicable in assessing diagnostic robustness since this is a real clinical situation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026



BenignCase | Malignantcase | Normalcase

### A. DataExploratoryAnalytics

Exploratory Data Analysis (EDA) was conducted to understand the characteristics of the CT image dataset prior to model training. This was analyzed by analyzing the distribution of benign versus malignant classes, analyzing changes in image resolution and intensity and visually analyzing representative CT slices.

The EDA demonstrated that there might be a class imbalance and the appearance of nodules could be highly varied, thus guiding the further preprocessing and augmentation efforts. Prior knowledge about the data was beneficial inensuring the probability of one-sided learning was reduced, and also, to avoid inappropriate generalization of the models.

The morphological variations of malignant, benign, and normal nodules were also to be understood by visual examination of representative CT slices. All these exploratory observations informed preprocessing decisions including normalization and augmentation strategies.

### B. DataCleaning

The quality and consistency of the CT images were also provided through data cleaning. First, the corrupted, duplicate, and low-quality images were eliminated. All images were also scaled to the same resolution to get the same input sizes to the model. Normalization of intensity was to be performed to normalize the pixel values as well as to enhance stability of the training.

To further enhance strength and the shortage of data, augmentation like rotation, horizontalflipping, zooming, and contrast adjustment were used. These changes increased the diversity of datasets and minimized on over-fitting because the model was able to learn invariant feature representations.

### C. DataSplittingSTesting

After preprocessing, the data set was divided into training, validation and test data sets based on a stratifiedsplittingapproachtomaintainthebalance of classes. Usually, the data was divided into 70%,20%,10% percent training and the remaining was used in testing. The training data was used to form a validation set that was used to monitor the performance of the model during training.

The testing set was never viewed in the training or validation process to provide an objective and credible information on the performance of the model.

### D. Modelling

A deep learning architecture was a Vision Transformer that was used to model lung nodule classification. Vision Transformers unlike other traditional CNNs focus on the localized characteristics, but image patches are treated as tokens and a self-attention mechanism is used to reveal the contextual relationship between globalfeatures.

The large range of nodule size and morphology was dealtwiththroughamulti-scalemodelingstrategy. Through the acquisition of representations at varying levels of spatial resolution, the model can achieve the ability to learn fine texture structure as well as global structure, important to malignancy detection.

### E. Model Architecture

The suggested Multi-Scale Vision transformer (MS- ViT) architecture is specially designed to extract both the local texture and global structural information of the lung CT images. The input image is divided into patches at different spatial resolutions to take into account the difference in lung nodule size and morphology. Every patch is then sent through a specific transformer branchthat is made up of multi-head self-attention and feed-forward network layers.

Using parallel transformer, scale-specific feature representations are learned independently by the parallel transformer branches. The results of the branches are then combined through a weighted fusion attention mechanism which assigns adaptively varying weights to features extracted at the various resolutions. The result of this combined representation is then transmitted to aclassification head to anticipate that a lung noduleis either benign or malignant. The multi-scale architecture brings in richness of features and classification robustness as compared with single- scale architectures.

### F. ModelEquation

Let the input CT image be denoted as X. The imageis divided into non-overlapping patches, which are linearly projected into an embedding space as:

$$Z_0=[X_p^1E;X_p^2E; \ldots;X_p^NE]+E_{pos}:$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

where $X_p$'represents the flattened image patches, E denotes the patch embedding matrix, and $E_{pos}$ represents positional embeddings.

The self-attention mechanism within each transformer layer is defined as:

$$Attention(Q,K,V)=softmax((QK^T)/\sqrt{d_k})V$$

where Q, K, and V are the query, key, and value matrices respectively, and $d_k$ is the dimensionality of the key vectors. The multi-scale featuresobtained from parallel transformer branches are fused using weighted attention and passed to the classification layer for final prediction.

### G. Trainingand Validation

The model was trained using the Adam optimizer and categorical cross-entropy loss and through several epochs using mini-batches of gradient descent. During the training process, accuracy in validation and loss were used to measure the performance of generalization.

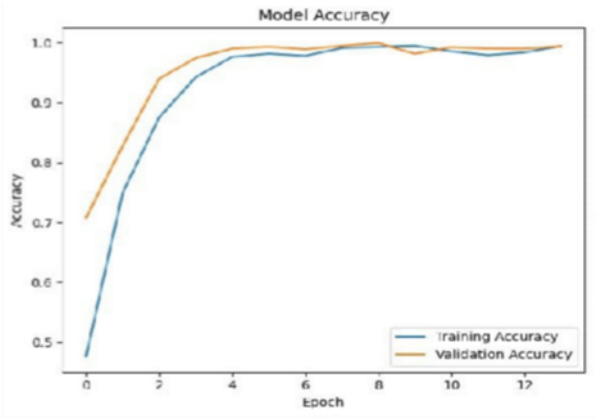



Figure3.1:VITModelTrainingmodelAccuracy

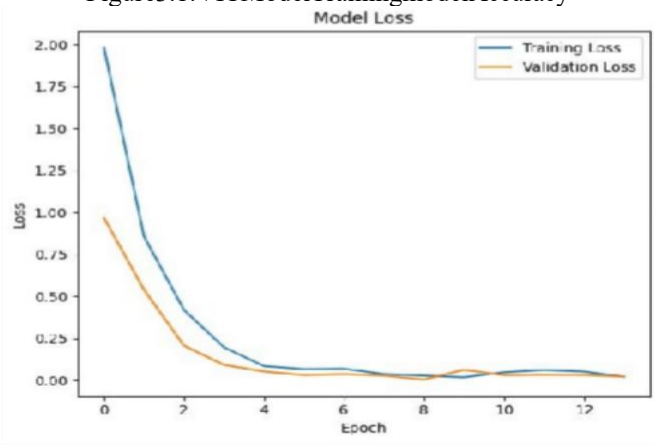



Figure3.2:VITModelTrainingModelLoss

The early stopping was used to reduce the overfitting risk when the validation performance stopped improving, which is why the convergenceis stable, and it can be trusted in its performance on unseen data.

## 2. RESULTSAND DISCUSSION

The experimental results support the fact that the MS ViT model is very effective when it comes to classifying the lung nodules as malignant, benign, and normal. The model smooths with little overfitting as demonstrated by the training and validation accuracy and loss curves. The confusion matrix shows that most of the malignant cases are mainly diagnosed correctly with very few being incorrectly diagnosed, which means the false- negative rate is low. To give an example, malignant nodules are highly sensitive, which is invaluable in avoiding missed oncology diagnosis in the clinical practice. Similarly, cases of benign and normal are well differentiated hence reducing unnecessary clinicalintervention.Theseresultsprovethatmulti- scale attention mechanism helps to increase reliability of the system and its diagnostic value.

As an illustration, over 99percent of the malignant nodules in the test dataset were rightly identified with a few false negatives. This has clinical implications because the untreated malignant cases can postpone treatment. The unnecessary follow up procedures were also minimized as benign and normal samples were correctly distinguished.

### A. AccuracyandLossGraphics

The accuracy of the validation and the loss curve demonstrate the learning behaviour of theproposed MSViT model with an increase in the number of epochs. Training and validation accuracy increase gradually, and the loss decreases continuously, which is an indication of a stable convergence of the model (as shown in the corresponding figure).

The goodness of extrapolation and the non- occurrence of overfitting is reflected by the close correspondence between training and validation curves. Such stability is explained by the use of multi-scale feature-extraction strategy and regularization methods used during the training.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

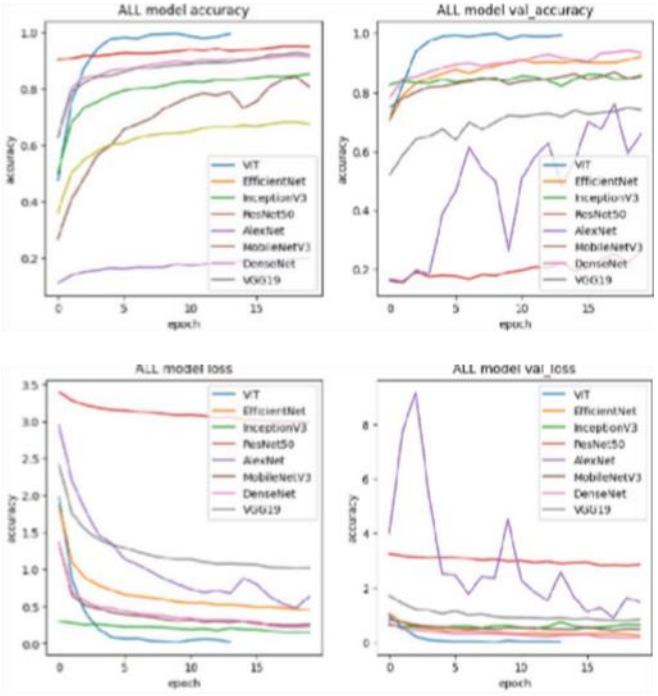



Figure4.1:Trainingandvalidationaccuracyand loss curves of the proposed MS-ViT model demonstrating stable convergence and minimal overfitting.

### B. ModelPerformanceComparison

The section of this paper is a comparative review of various deep-learning models that have been tested to classify lung nodules into malignant and benign. Each model is evaluated by the accuracy of the validation obtained, to determine the best architecture.

Table 4.1: Validation Accuracy Comparison of Different Models

| Model | Architecture Type | Validation Accuracy |
|---|---|---|
| CustomCNN | CNN | 99.64% |
| VGG19 | CNN(Deep) | 80.56% |
| ResNet50 | CNN(Residual) | 85.70% |
| InceptionV3 | CNN (Multi-scale) | 88.92% |
| EfficientNetB0 | CNN (Optimized) | 91.93% |
| VIT | Transformer | 99.95% |

### C. ExplanationofResults

As the experimental findings show, the proposed Multi-Scale Vision Transformer (MS-ViT) model has a higher classification in the ability to separate malignant, benign, and normal nodules of lung cancer on the basis of CT images. The MSViT model, as shown by the validation accuracy curve and loss curve, has a steady convergence, and a small difference between training and validation, which suggests that it has good generalization ability. The peak validation rate of the model is 99.9 % , whichis superior to any classic CNN-based architectures tested in this paper.

The effectiveness of the proposed strategy is mostly explained by the fact that the model implements a multi-scale patch-embedding strategy that allows resolving a fine-grained texture detail and global structural attribute of lung nodules. The transformer architecture is an architectural design that uses a self-attention system to capture long-range spatial interactions in contrast to conventional convolutional neural networks that make use of localized receptive fields. This is significant because, this ability significantly enhances the discrimination between similarmalign and benign nodules that are visually similar and is a significant challenge in the lung cancer diagnosis. The outcomes made it very clear that the useofmulti-patchresolutionsprovidesbetter

feature representation than the use of single-scale methods, thus improving the overall diagnostic accuracy.

### D. ConfusionMatrixGraph

The confusion matrix gives the detailed results of the classification into the true positives, true negatives, false positives and false negatives. The MS-ViT model as shown in the figure is able to classify most of the benign and malignant samples correctly.

In medical diagnosis, a high true-positive rate is of special importance because it minimizes thechances of false negative diagnosis in cases of malignancy. On the same note, the false positivesare few thus reducing redundant clinical interventions.

The confusion matrix demonstrates the high sensitivity and specificity especially in cases of malignancy, which adds to the credibility of the proposed method in clinical screening scenarios.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

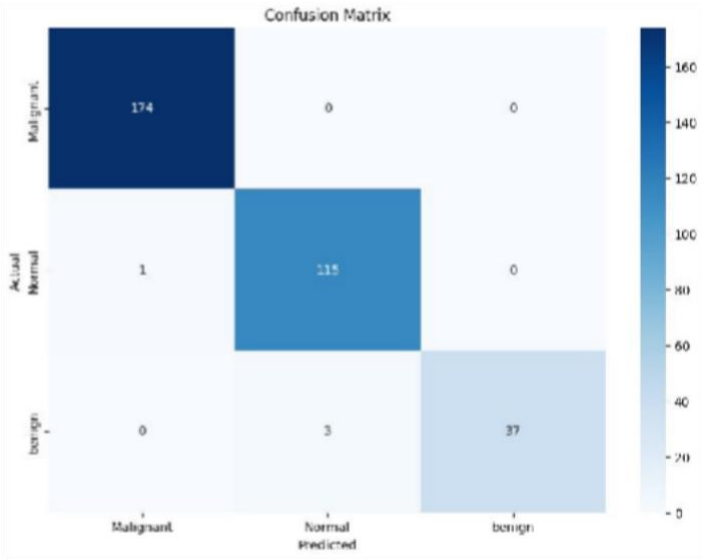



Figure4.2–ConfusionmatrixoftheproposedMS- ViT model for lung nodule classification

### E. Analysisofthe Results

The results further indicate the strength andvalidity of the suggested MS -ViT framework. This high performance can be explained by integrationof parallel transformer branches with different spatial scale as well as weighted fusion attention mechanism that adaptively fuses multi-scale features.

Also, attention-based interpretabilitymethods demonstrate that the model concentrates on clinicallysignificantareasoflungnodules,including irregular edges and non-homogeneous interior structures. This increases transparency and confidence in predictions of the model.

All in all, the findings prove the effectiveness of the proposed method compared to traditional CNN-basedandsingle-scaletransformermodels,whichis whyitcanbeasuccessfulsolutiontotheproblemof computer-aided lung cancer diagnosis.

### F. ComparativeDiscussionwithExisting CNN Models:

The proposed MS -ViT model was tested qualitatively and compared to the traditional CNN - based models, i. e., VGG, ResNet, and DenseNet, as already presented in the literature. The MS-ViT architecture uses self-attention throughout thebody and uses multi-scale feature extraction as opposed to CNN models that use localized receptive fields, which provides the model with better contextual understanding of lung nodules. It is also much more stable, more generalized and fewer false-negative predictions, thus showing that the method is superior to the old single-scale, convolution-based models in the area of nodules in the lungs.

## 4. CONCLUSION&FUTURE WORK

This paper described a Multi-Scale Vision Transformer (MS -ViT) architecture to classify lung nodules based on computed-tomography scan images. The proposed method combines multi- resolution patch embeddings with a transformer- based state of self-attention, therefore, successfully extracting fine-grained texture features and whole structural context. The experimental findings show that the MS-ViT model can get high accuracy in classification with the stability of training behavior compared to the traditional CNN-based and single- scale transformer models.

Parallel transformer branches and weighted fusion attention allow the representation of features of nodules of different sizes and morphology. Moreover,attention-basedinterpretabilitymethods improve transparency by pointing out clinically significantareas,whichmakesitmoretrustworthy in the predictions made by the model. These results validate the appropriateness of the transformer- based architectures to computer aided diagnosis of lung cancer and how such systems can help radiologists to detect it early and accurately.

The achieved accuracy of 99.9% demonstrates the effectiveness of transformer-based multi-scale representations for medical image analysis.

Although the results are promising, there are still a number of avenues on which future research can be done. The extension of the proposed framework to the three-dimensional (3D) CT volumes to further harness the spatial continuity between the slices will be the work in the future. The inclusion ofmulti-view and multi-phase CT data would also increase the accuracy of the diagnosis. Also, incorporating radiomic features into transformer- based representations and testing the model on the multi-institutional level will enhance both robustness and generalization, which will facilitate the application of the model into clinical practice.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

## REFERENCES

[1] M. K. Faizi et al., "Deep learning-based lung cancer classification of CT images," *BMC Cancer*, vol. 25, 2025. Available: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-025-14320-8

[2] L. Wang et al., "Deep learning techniques to diagnose lung cancer," *Frontiers in Oncology*, 2023. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9688236/

[3] M. A. Thanoon et al., "A review of deep learning techniquesforlungcancer,"*Cancers*,2023. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10453592/

[4] N. Nasrullah et al., "Automated lung nodule detection and classification using deep learning techniques," *Sensors*, vol. 19, no. 17, 2019. Available: https://www.mdpi.com/1424-8220/19/17/3722

[5] W. Hendrix et al., "Deep learning for benign and malignant pulmonary nodules," *Nature Communications*, 2023. Available: https://www.nature.com/articles/s43856-023-00388-5

[6] Z. UrRehman et al., "Effective lung nodule detectionusingdeepCNNwithdualattention,"

[7] *Scientific Reports*, 2024. Available: https://www.nature.com/articles/s41598-024-51833-x

[8] R. Durgam et al., "Enhancing lung cancer detection with Vision Transformers," *Scientific Reports*, 2025. Available: https://www.nature.com/articles/s41598-025-00516-2

[9] D.Riquelmeetal.,"Deeplearningforlungcancer nodulesdetectionandclassification:Asurvey,"

[10] *Diagnostics*, 2020. Available: https://www.mdpi.com/2673-2688/1/1/3

[11] K. Abdullah et al., "Deep learning techniques for lungcancerdiagnosis:Acomprehensivereview,"

[12] *Information*, vol. 16, no. 6, 2025. Available: https://www.mdpi.com/2078-2489/16/6/451

[13] H. Mkindu et al., "Computer-aided diagnosis of lung cancer using Vision Transformers,"

[14] *BiomedicalSignalProcessingandControl*,2023. Available: https://www.sciencedirect.com/science/article/abs/pii/S1746809423002999

[15] J. Debnath et al., "A hybrid Vision Transformer for lung cancer detection," *Artificial Intelligence in Medicine*, 2025. Available: https://www.sciencedirect.com/science/article/pii/S2352914825000577

[16] M. Q. Shatnawi, "Deep-learning-based CT lung cancer diagnosis," *International Biomedical Engineering*, 2025. Available: https://www.sciencedirect.com/science/article/pii/S2666521224000553

[17] Z. Gao et al., "A lung CT vision foundation model for disease diagnosis," *Nature Communications*, 2025. Available: https://www.nature.com/articles/s41467-025-66620-z

[18] A. Kumar, "Vision Transformer model for lung cancer detection," *SN Computer Science*, 2024. Available: https://link.springer.com/article/10.1007/s42979-024-03120-9

[19] T. Z. Li et al., "Time-distance Vision Transformers for lung cancer classification," *arXiv preprint*, 2022. Available: https://arxiv.org/abs/2209.01676

[20] R.Sunetal.,"SwinTransformerforlungcancer classification," *Electronics*, vol. 12, no. 4, 2023. Available: https://www.mdpi.com/2079-9292/12/4/1024

[21] R.Pandianetal.,"Detectionandclassification oflungcancerusingCNNanddeeplearning,"

[22] *MaterialsToday:Proceedings*,2022.Available: https://www.sciencedirect.com/science/article/pii/S2665917422002227

[23] T. Gulsoy et al., "Diagnosis of lung cancer based onCTscansusingVisionTransformers,"

[24] *ResearchGate Preprint*, 2023. Available: https://www.researchgate.net/publication/378014908

[25] A.Naiketal.,"LungnoduleclassificationonCT: A comprehensive survey," *Wireless Personal Communications*, 2021. Available: https://link.springer.com/article/10.1007/s11277-020-07732-1

[26] L.J.Crastaetal.,"Anew3Dlungcancer detection architecture using deep learning,"

[27] *Machine Learning with Applications*, 2024. Available: https://www.sciencedirect.com/science/article/pii/S2772442524000182

[28] A.Paletal.,"ViT-DCNNhybridmodelforenhanced lungand colon cancer detection,"

[29] *ComputersinBiologyandMedicine*,2025.Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12468260/