SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

ENHANCING LUNG NODULE DETECTION AND CLASSIFICATION USING VISION TRANSFORMERS IN MEDICAL IMAGING

Muhammad Mashood Khan[1], Hafza Eman[*2], Ishtiaque Mahmood[3], Abdullah Danish[4], Marium Mumtaz[5]

[1]*University of Engineering and Technology, Taxila*
[*2]*HITEC University, Taxila*
[3]*Oman College of Management and Technology, Oman*
[4]*University of Engineering and Technology, Taxila*
[5]*HITEC University, Taxila*

[*2]hafza.eman@hitecuni.edu.pk

DOI: https://doi.org/10.5281/zenodo.20609220

Keywords
MedSAM, MobileViT, Lung Nodule Detection, Lung Nodule Classification, Lung Cancer, Self-attention, Transformer, CNN

Article History
Received: 07 April 2026
Accepted: 19 May 2026
Published: 09 June 2026

Corresponding Author: *
Hafza Eman

Abstract

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, primarily due to late-stage diagnosis and the difficulty of accurately identifying pulmonary nodules in early stages. Computed Tomography (CT) imaging plays a vital role in lung cancer screening; however, manual interpretation of CT scans is time-consuming, prone to inter-observer variability, and often affected by the subtle and highly variable nature of lung nodules. To address these challenges, this study proposes an automated lung nodule detection and classification framework based on deep learning techniques. The proposed approach integrates MedSAM based segmentation with a MobileViT based classification model to improve both accuracy and computational efficiency. Initially, lung nodules are segmented from CT images using MedSAM. The segmented nodules are then passed to a MobileViT network, which combines convolutional layers for local feature extraction with transformer-based self-attention mechanisms for capturing global contextual relationships. This hybrid design enables the model to effectively learn both fine-grained morphological features and long-range dependencies within nodule regions. The framework is evaluated on the LIDC-IDRI dataset and achieves strong performance with a training accuracy of 95.58%, validation accuracy of 92.13%, and test accuracy of 91.30%. Experimental results demonstrate that the proposed method provides stable learning behavior, reduced misclassification rates, and balanced performance across benign and malignant classes. The integration of segmentation and classification further improves robustness by focusing the model on clinically relevant regions and reducing background noise.

INTRODUCTION

Lung cancer is one of the most common and deadliest types of cancer worldwide. It causes a large number of cancer-related deaths every year and remains a major public health challenge [1]. One of the most effective ways to improve patient survival is through early detection. Pulmonary nodules, which are small masses of tissue found

in the lungs, can be an early sign of lung cancer. Detecting and classifying these nodules at an early stage can help doctors begin treatment sooner and improve patient outcomes [2].

Computed Tomography (CT) scans are widely used to detect pulmonary nodules because they provide detailed images of the lungs. However, analyzing CT scans manually can be difficult and time-consuming. Radiologists often need to examine hundreds of images for a single patient, which increases the risk of missing small nodules. In addition, pulmonary nodules can vary greatly in size, shape, texture, and location, making accurate diagnosis even more challenging. Differences in interpretation among radiologists may also lead to inconsistent results and diagnostic errors [3].

To overcome these challenges, researchers have developed Computer-Aided Detection (CAD) systems based on deep learning techniques. These systems can automatically analyze medical images and help radiologists identify suspicious nodules more accurately and efficiently. Convolutional Neural Networks (CNNs) have been widely used for this purpose and have shown good performance in medical image analysis [4]. However, CNN-based models mainly focus on local image features and may not always capture the broader context of the image, which is important for accurate classification.

Recently, Vision Transformers (ViTs) have gained significant attention in medical imaging [5]. Unlike traditional CNNs, ViTs use a self-attention mechanism that can capture both local and global information from an image. This allows them to better understand complex patterns and relationships within medical images. As a result, ViTs have shown promising performance in various image classification and detection tasks, including lung nodule analysis [6]. Despite these advantages, existing ViT-based methods still face several challenges. Their performance may decrease when applied to images from different hospitals, scanners [7], or patient groups. Furthermore, the large variation in pulmonary nodules makes accurate detection and classification difficult [8]. These limitations highlight the need for more robust and reliable models that can perform consistently across diverse clinical settings [9].

Therefore, this research aims to develop an enhanced Vision Transformer-based framework for lung nodule detection and classification. The proposed approach seeks to improve model robustness, increase classification accuracy, and reduce false positive and false negative predictions. By providing more reliable support for radiologists, the proposed system has the potential to assist in earlier diagnosis of lung cancer and contribute to better patient care and clinical decision-making.

Literature Review

Recent advances in deep learning have significantly improved the automated detection and classification of pulmonary nodules from Computed Tomography (CT) images. Traditional Convolutional Neural Network (CNN)-based methods have demonstrated strong performance in extracting local image features; however, they often struggle to capture long-range spatial dependencies and global contextual information. To address these limitations, researchers have increasingly adopted transformer-based architectures, particularly Vision Transformers (ViTs), which utilize self-attention mechanisms to learn both local and global image representations.

Miao et al. [10] proposed a transformer-based framework for differentiating subtle ground-glass nodules (GGNs) using volumetric CT scans. Their approach learns three-dimensional asymmetry features through a self-attention mechanism that captures long-range spatial relationships within lung volumes. Experimental results demonstrated improved discrimination between benign and malignant GGNs compared to conventional CNN approaches. However, the model relies heavily on accurate initial nodule localization and was evaluated on a relatively limited dataset, raising concerns regarding generalization to larger and more diverse clinical populations.

To improve volumetric feature learning, Oumlaz et al. [11] introduced ARSGNet, a hybrid architecture combining EfficientNet feature extraction, transformer attention modules, and

adaptive slice grouping mechanisms. The proposed framework effectively models cross-slice relationships and preserves contextual information across CT volumes. Their results showed improved sensitivity for small nodules and better robustness against image variations. Despite these advantages, the architecture consists of multiple interconnected modules, making implementation and parameter optimization considerably more complex.

Mahmoud et al. [12] proposed a lightweight dual-output Vision Transformer architecture designed for pulmonary nodule classification and localization. In addition to predicting nodule classes, the model generates attention maps that improve interpretability for radiologists. The study demonstrated competitive classification performance while maintaining low computational requirements, making it suitable for deployment in resource-constrained healthcare environments. However, the reduced model size may limit its ability to capture highly complex nodule characteristics compared with larger transformer ensembles.

A significant advancement in volumetric CT analysis was introduced by Cao et al. [13], who developed a three-dimensional multifaceted attention encoder capable of extracting context-aware nodule representations. By integrating multiple attention mechanisms, the model captures fine-grained spatial interactions between neighboring anatomical structures and pulmonary nodules. Experimental evaluations showed substantial improvements in classification accuracy, particularly for irregular and heterogeneous nodules. Nevertheless, the computational and memory requirements of volumetric attention operations remain a major limitation for practical clinical deployment.

To address segmentation challenges, Hu et al. [14] proposed a dual-encoding fusion network that combines CNN-based local feature extraction with transformer-based global contextual learning. This hybrid approach effectively preserves nodule boundaries while maintaining awareness of surrounding anatomical structures. The method achieved superior segmentation accuracy for atypical pulmonary nodules, especially those with irregular shapes. However, the study primarily focused on segmentation performance and did not provide a complete end-to-end framework for pulmonary nodule detection and classification.

Recognizing the complementary strengths of CNNs and transformers, Ma et al. [15] introduced TiCNet, a hybrid architecture that integrates transformer blocks into convolutional networks. The model simultaneously captures local texture information and global contextual features, leading to improved detection and classification performance. Results on public lung nodule datasets demonstrated enhanced robustness across nodules of varying sizes and appearances. Despite these improvements, the increased network depth results in longer training times and greater computational demands.

Pal et al. [16] further explored hybrid architectures by combining deformable convolutional layers with Vision Transformer attention mechanisms. Deformable convolutions enable adaptive feature extraction from nodules with diverse shapes and textures, while transformer modules capture long-range dependencies. Their results showed increased sensitivity in identifying subtle structural abnormalities. However, the model exhibited a tendency toward overfitting when trained on relatively small datasets, highlighting the need for larger annotated medical image repositories.

Swin Transformer-based approaches have recently gained considerable attention due to their hierarchical feature extraction capabilities. Wu et al. [17] employed a Swin Transformer architecture for benign and malignant pulmonary nodule classification. By utilizing shifted-window attention mechanisms, the model effectively captured multi-scale image features and achieved superior classification performance compared to conventional CNN models. Nevertheless, the effectiveness of Swin Transformers depends heavily on large-scale pretraining and sufficient training data.

Similarly, Wen et al. [18] applied Swin Transformer networks for predicting pathological subtypes of pure ground-glass nodules. Their

framework leveraged contextual image information to improve subtype classification accuracy and demonstrated strong agreement with expert radiologists. However, the model's performance remains dependent on the quality and diversity of training datasets.

Recent studies have also investigated optimization strategies for Vision Transformer training. Ko et al. [19] evaluated multiple optimization techniques for transformer-based lung disease classification systems. Their findings revealed that optimizer selection significantly influences convergence speed, model stability, and final classification performance. While their conclusions provide valuable insights for transformer training, the study focused primarily on chest radiographs rather than volumetric CT scans, limiting direct applicability to pulmonary nodule analysis.

Mkindu et al. [20] proposed a Vision Transformer-based Computer-Aided Detection (CAD) system enhanced with Bayesian optimization for automated hyperparameter tuning. The optimization process improved detection sensitivity while reducing false-positive rates. Although the framework achieved promising results, the additional optimization stage increased overall computational complexity and training duration.

To improve feature representation, Du et al. [21] integrated squeeze-and-excitation channel attention modules into a Vision Transformer framework. This enhancement enabled the model to emphasize diagnostically relevant image channels and suppress less informative features. Experimental results demonstrated improved classification performance, particularly for subtle pulmonary nodules. However, broader validation on larger multi-center datasets is still required.

A notable recent contribution was presented by Wang et al. [22], who proposed a two-stage pulmonary nodule detection pipeline combining U-Net segmentation, YOLOv8 candidate detection, and Swin Transformer refinement. The framework significantly reduced false positives while maintaining high detection sensitivity. Although the sequential design improved overall performance, errors generated during early segmentation stages could propagate throughout the pipeline and negatively affect final predictions.

Sun et al. [23] introduced a dual-fusion transformer architecture incorporating gated fusion mechanisms to preserve high-resolution image details. The model demonstrated improved detection accuracy for very small pulmonary nodules, which are often missed by conventional systems. Nevertheless, the additional fusion modules increased inference latency and computational complexity.

Recent studies have also explored the integration of transformer architectures with advanced multimodal and clinical information to improve pulmonary nodule analysis. Eman et al. [24] proposed EMeRALDS, an end-to-end computer-aided diagnosis framework that combines vision-language models with CT image analysis. The framework utilizes a modified Segment Anything Model (SAM2) guided by CLIP-based text prompts for pulmonary nodule detection and further incorporates radiomic and clinical information for classification. By integrating imaging features with electronic medical record information, the system improves diagnostic support and enhances contextual understanding of pulmonary nodules. However, the framework introduces additional complexity due to its dependence on multimodal data sources and the availability of structured clinical records.

To address challenges related to data privacy and limited access to centralized medical datasets, Turjya and Fawakherji [25] proposed a federated hybrid Transformer–U-Net architecture for pulmonary nodule segmentation. Their framework combines transformer-based global contextual learning with U-Net feature extraction while enabling collaborative model training across multiple institutions without sharing patient data. The study demonstrated improved segmentation performance for small and low-contrast nodules while preserving data privacy. Although the federated approach increases model generalization across diverse datasets, communication overhead and training synchronization between institutions remain significant challenges.

Eman et al. [26] proposed an advanced deep learning framework for early detection and classification of lung cancer by integrating the Segment Anything Model 2 (SAM2) with DenseNet architectures. In this study, SAM2 is employed for accurate segmentation of pulmonary nodules from CT images using prompt-based learning, while DenseNet is utilized for classification into benign and malignant categories. The proposed hybrid approach demonstrates strong performance, achieving high segmentation accuracy and improved classification. However, despite its promising results, the approach may face limitations in generalization when applied to multi-center datasets and real-world clinical environments due to variability in imaging protocols and patient populations. This indicates the need for more robust and domain-adaptive models for reliable lung cancer detection in diverse clinical settings.

The reviewed studies demonstrate the growing effectiveness of transformer-based and hybrid CNN-transformer architectures for pulmonary nodule detection and classification. While Vision Transformers successfully address many limitations of traditional CNNs by capturing global contextual information, challenges remain regarding computational efficiency, generalization across different datasets, limited annotated medical data, and robustness in real-world clinical environments. These limitations highlight the need for an enhanced Vision Transformer-based framework capable of achieving high accuracy while maintaining computational efficiency and strong generalization performance across diverse patient populations.

Objectives:

1. To develop an automated lung nodule segmentation approach using MedSAM to accurately extract nodule regions from CT images.
2. To design a robust classification model based on MobileViT that combines convolutional feature extraction with transformer-based self-attention.
3. To improve the performance of the proposed MedSAM and MobileViT framework on the LIDC-IDRI dataset in terms of accuracy, precision, recall, and F1-score.

Materials and Methods

Dataset Description

In this work we used the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset [3]. It is shared publicly and used extensively in medical imaging research. The dataset, obtained from the original LIDC-IDRI database of annotated thoracic CT scans is available on Kaggle.

The LIDC-IDRI dataset includes 1,018 thoracic computed tomography (CT) scans acquired with different imaging acquisition parameters from different clinical centers. Each CT scan is annotated with a detailed lesion map, which is obtained from a structured two-phase reading by four thoracic radiologists (see Figure 1). For this, the scans were independently reviewed by radiologists who categorized the lesions into nodules > 3 mm, nodules < 3 mm, and non-nodules > 3mm, supporting detection and characterization studies. Figure 1 presents a sample CT scan image and Masks.

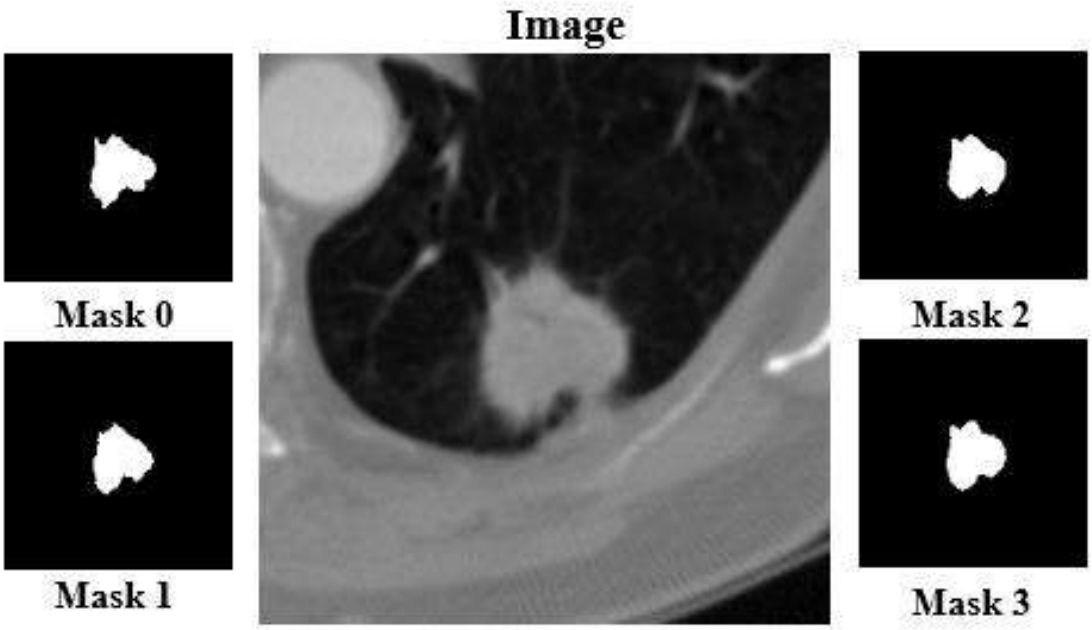

Fig 1: Sample image and Segmentation masks from dataset

Proposed Methodology

In our work, we proposed a framework for lung nodule analysis including segmentation, feature extraction and classification. MedSAM was first used to segment lung nodules in thoracic CT images. Then the segmented nodule regions are used as an input to MobileViT for feature extraction and classification. This approach leveraged the segmentation ability of MedSAM and the transformer learning ability of MobileViT to achieve accurate and computationally efficient lung nodule detection and classification. The detailed proposed methodology is depicted in Figure 2.

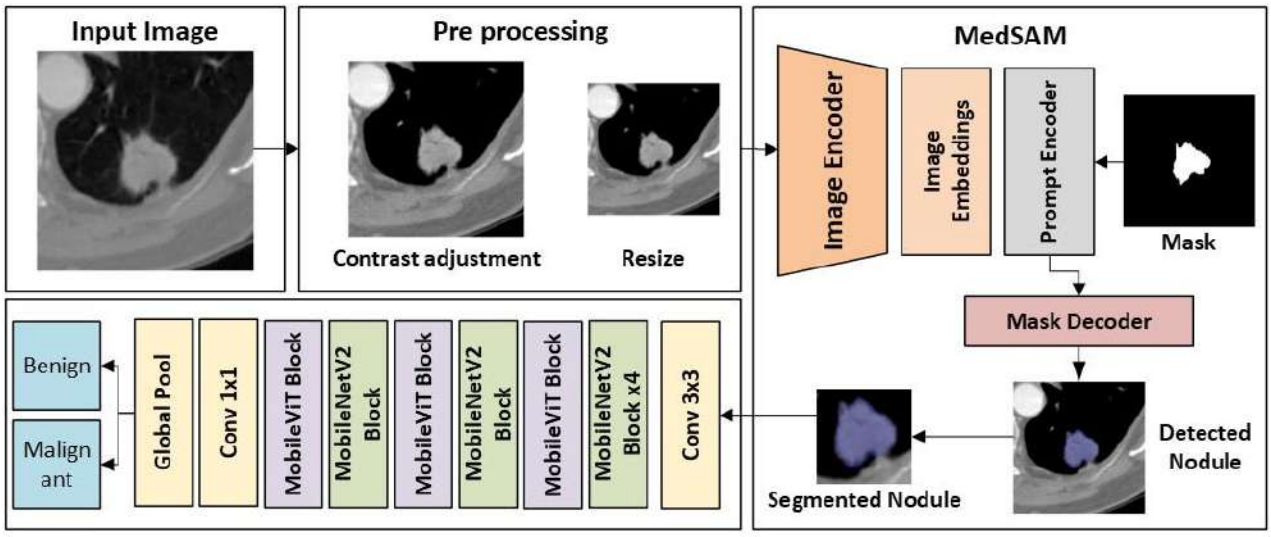

Fig 2: Proposed Methodology for Lung Nodule Detection and Classification

Preprocessing

All input CT images are subjected to a standardized pre-processing pipeline to improve the image quality and to make them compatible with the MedSAM framework. This step is needed in order to improve contrast, normalize spatial dimensions and reduce variability across scans. The first step is to adjust the contrast to improve the visibility of lung structures and possible nodules. Medical CT images often suffer from low contrast or inconsistent intensity distribution which can be attributed to different acquisition protocols and patient-specific factors. This is achieved by normalizing intensity values and enhancing contrast to make subtle nodular regions more distinguishable from surrounding tissue. All images are resampled to a fixed spatial resolution after contrast enhancement. This ensures a consistent input size for further processing and enables fast batch training and inference in the deep learning architecture. Resizing preserves anatomical structure and reduces computational complexity.

Nodule Segmentation

We used MedSAM [27] for the task of lung nodule segmentation. MedSAM is the extension of Segment Anything Model (SAM) [28] to The Medical Domain. MedSAM is designed for medical image analysis tasks, while keeping the wonderful generalization ability of SAM. The prompt-driven segmentation framework allows its application to lung nodule analysis, where nodules often exhibit variability in size, shape, texture, and intensity across patients and imaging conditions.

The segmentation process is guided by the expert annotations provided in the LIDC-IDRI dataset where each lung nodule is independently annotated by four experienced radiologists [29]. These annotations are binary masks indicating the presence or absence of nodular regions. However, the subjective nature of medical interpretation leads to frequent differences among radiologists in defining the exact boundaries of nodules. This causes inter-observer variability which needs to be mitigated for better reliability. Therefore, a consensus-based fusion strategy is adopted. More specifically, the majority voting mechanism is employed on the pixel level such that a pixel is labeled as a part of a lung nodule only when labeled as a lung nodule by at least three radiologists, otherwise it is labeled as background. This consensus formulation effectively reduces the noise introduced by the annotation process and ensures that only high-confidence regions are kept for further processing.

The obtained consensus mask is then used as spatial prompt to MedSAM. This guides the model towards the clinically relevant regions during segmentation by providing the CT scan image together with the consensus mask that is validated by a radiologist. In the MedSAM architecture, the image encoder first processes the input image to obtain meaningful feature embeddings, and the prompt encoder combines spatial information extracted from the consensus mask. These complementary representations are then merged in the mask decoder, which allows the model to refine the initial annotations and precisely annotate lung nodules.

This prompt-driven refinement mechanism enables MedSAM to generate segmentation outputs that precisely capture the spatial extent and boundaries of nodules while retaining important anatomical context. The generated masks are robust to variations in imaging characteristics and nodule appearance and thus perform consistently across different cases. Thus, the segmented nodule regions serve as a reliable input for the subsequent steps of feature extraction and classification, enabling the downstream models to concentrate exclusively on the clinically relevant regions.

Nodule Classification

A MobileViT [30] based architecture is used to classify lung nodules after the segmentation stage. MobileViT is designed to learn discriminative representations in a computationally efficient manner. The segmented nodule regions provided by MedSAM are then fed into the classification network. This approach forces the model to pay attention only to clinically important structures, by isolating the region of interest before

classification. This reduces the influence of surrounding lung parenchyma, vessels or background artifacts that could otherwise add noise to the learning process.

Then the segmented nodule images are sent into the MobileViT network, which combines the advantages of convolutional neural networks and transformer-based learning. The convolutional layers are responsible for extracting fine-grained local features such as edges, textures, and shape irregularities that are crucial for characterizing nodular morphology. These layers are critical for detecting subtle visual cues that distinguish benign and malignant nodules. Simultaneously, the transformer-based components learn long-range dependencies in the segmented region via self-attention mechanisms. This enables the network to learn wider contextual relationships within the nodule, such as internal heterogeneity or structural asymmetry, which are key indicators in clinical assessment.

In MobileViT, as the feature maps progress through the network, the low-level spatial information is progressively converted into high-level semantic representations. The hybrid architecture allows the concurrent learning of fine morphological patterns and global contextual properties, hence providing a more informative feature representation than traditional CNN approaches. This learned representation is then aggregated by the classification head, which consists of global pooling and fully connected layers, to synthesize the extracted information into a compact descriptor.

The last output layer generates class probabilities for clinically relevant categories to differentiate between benign and malignant nodules. The network optimizes feature extraction and classification simultaneously in training, and thus can learn robust patterns with pathological variations. This end-to-end learning strategy improves the discriminative capacity of the model and increases the classification accuracy.

Results and Discussion

The proposed lung nodule detection and classification framework was trained and evaluated on the LIDC-IDRI dataset on Google Colab with GPU acceleration, which enabled efficient training of the deep learning models and reduced computational time. Standard accuracy metrics were used to evaluate the model's performance on the training, validation and test sets, in order to assess its learning capability and generalization performance.

The experimental results indicate that the proposed method achieves a training accuracy of 95.58%, which means the model can learn discriminative patterns from the segmented lung nodules effectively. The validation accuracy of 92.13% suggests that the model is generalizing well during training, and does not overfit the training data too much. Moreover, the model achieves a test accuracy of 91.30%, showing its robustness and reliability when tested on unseen data as shown in Figure 3. The small difference between training, validation and test accuracies indicates stable learning behavior and good generalization ability.

A number of factors can explain the strong performance. First, we can accurately locate lung nodules by segmenting nodules using MedSAM, which enables the classification model to only focus on areas of clinical interest and reduce interference from the background. Second, the MobileViT architecture, combining convolutional feature extraction with transformer-based global attention, can effectively capture both local morphological features and long-range contextual information for accurate lung nodule classification. The lightweight design also lends itself to deployment in resource-constrained environments. Finally, training the model in a GPU environment helped to optimize and converge efficiently.

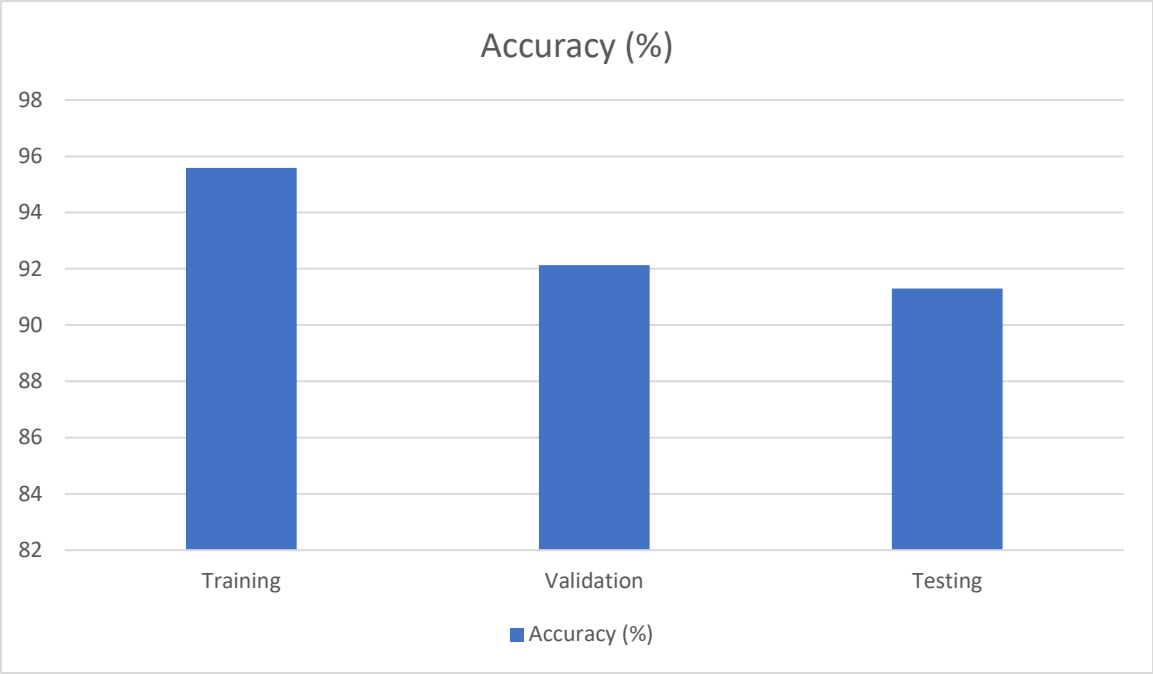

Fig 3: Training, Validation and Testing Accuracy of the Proposed Model

The detailed evaluation on test set further shows the effectiveness of the proposed lung nodule classification framework, as shown in Table 1. The model converged stably and achieved a test loss of 0.3861 and test accuracy of 91.30%, showing its stable convergence and reliable predictive performance on unseen data. Classification report There is a good balanced performance on both the classes. For benign nodules, the model achieved a precision, recall and F1-score of 0.92, indicating its ability to correctly identify non-malignant cases with limited false positives and false negatives. The model achieved a precision, recall and F1-score of 0.91 for malignant nodules, indicating its robustness in identifying clinically important malignant cases. The macro and the weighted average score for precision, recall and F1-score are 0.91. This denotes that the performance is balanced across all the classes and the model is not biased towards any specific category. These results show the robustness and clinical applicability of the proposed MedSAM and MobileViT framework for accurate lung nodule classification.

Table 1: Class wise classification performance of the model

Class	Precision	Recall	F1-Score
BENIGN	0.92	0.92	0.92
MALIGNANT	0.91	0.91	0.91
Accuracy	-	-	0.91
Macro Avg	0.91	0.91	0.91
Weighted Avg	0.91	0.91	0.91

The accuracy curves of training and validation shown in Figure 4 reflect the learning behavior and generalization capability of the proposed framework for lung nodule classification. With the advancement of training, the training accuracy steadily increases, rising from about 89% in the initial epochs to roughly 96% by the conclusion of the training process, indicating that

the model successfully acquires discriminative features from the segmented nodule regions. Likewise, the validation accuracy is progressively increasing, indicating that the model is able to generalize well on the new data that it has not seen before. The training and validation curves are separated by a small gap but the gap is fairly stable over the epochs. This indicates that there is not much overfitting and the model is learning well. The gradual convergence of both curves shows that the hybrid architecture is successful in capturing both local morphological and global contextual features required to distinguish benign from malignant nodules. The accuracy trends validate that the fusion of MedSAM-based segmentation and MobileViT classification facilitates robust feature learning and aids in reliable performance for lung nodule classification.

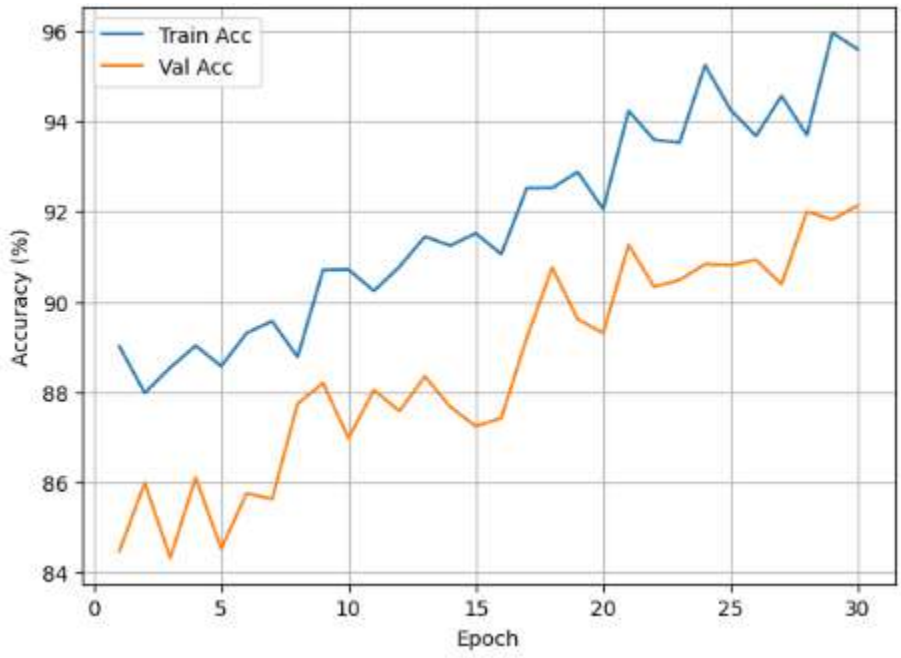

Fig 4: Training and Validation Accuracy Curves

The confusion matrix in Figure 5 also displays the classification performance of the proposed MobileViT-based model on the test dataset. In case of benign nodules, 91.67% of samples were correctly classified and 8.33% were misclassified as malignant which shows a low false-positive rate. For malignant nodules, the model achieved a correct classification rate of 90.91%, while 9.09% of malignant cases were misclassified as benign. The high values along the diagonal of the confusion matrix show the strong discriminative ability of the model and balanced performance across both classes. Importantly, the low misclassification rate of malignant nodules shows the clinical relevance of the proposed approach, as it is crucial to correctly identify malignant cases for early diagnosis of lung cancer. Overall, the confusion matrix confirms that the model maintains a good balance between sensitivity and specificity, which supports its utility in lung nodule classification tasks.

Fig 5: Confusion Matrix showing classification performance

The analysis of misclassification metrics of the MobileViT model presented in Figure 6 indicates a strong and well-balanced diagnostic performance across the target classes. Based on the distribution of total errors, the model achieved an overall classification success rate of 91.3% and a total misclassification rate of only 8.7%. Furthermore, we show the robustness of the model by providing a detailed analysis of the class-wise performance, with an error rate of 8.3% for Benign and 9.1% for Malignant cases. The minute difference of 0.8% between the two classes is very crucial as it indicates that the model is not undergoing any biased learning pattern or preference for one class over the other. This parity is crucial for medical imaging tasks, where both diagnostic classes need to be recognized with almost the same trustworthiness to reduce the risk of systematic errors, such as a high number of false negatives in malignant cases that could compromise clinical outcomes.

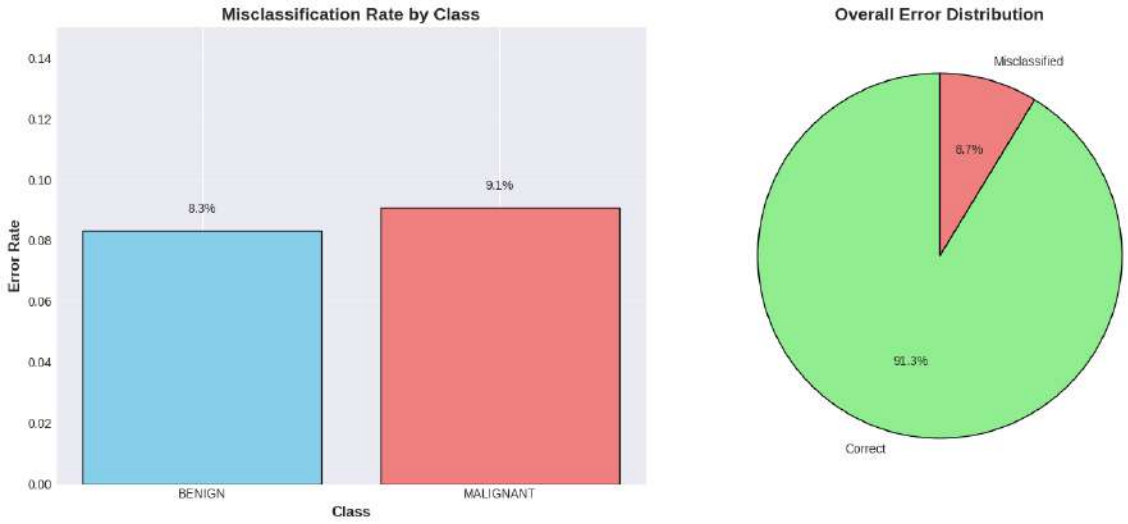

Fig 6: Misclassification rate by class and Overall Error Distribution chart

Conclusion:

In this work, we have proposed an integrated framework for automated lung nodule analysis, which combines segmentation with efficient classification to facilitate computer-aided diagnosis. Figure 7 shows the training and validation loss. The approach used MedSAM for accurate prompt guided segmentation of lung nodules and then classification using a MobileViT architecture. Expert annotations from the LIDC-IDRI dataset are used in the framework to ensure that nodules are reliably localized before being classified. The segmentation stage allows for accurate segmentation of lung nodules. The MobileViT network merges convolutional feature extraction with transformer based contextual learning, allowing for robust identification of subtle morphological differences between benign and malignant nodules. The experimental results show that the hybrid design can achieve stable learning behavior, strong inference ability and better generalization, which verifies the effectiveness of the hybrid design. The proposed pipeline provides a robust and computationally efficient framework for the analysis of lung nodules. Its ability to combine spatial precision with discriminative feature learning makes it well suited for real world deployment in computer aided diagnosis systems. This framework has the potential to assist radiologists in early detection and clinical decision making, ultimately leading to improved diagnostic accuracy and patient outcomes.

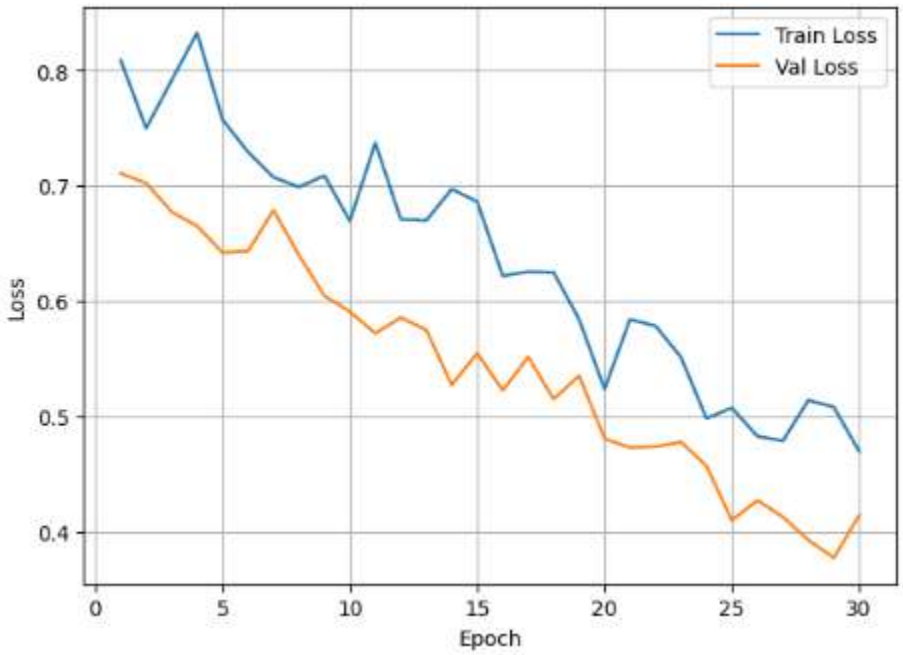

Fig 7: Training and Validation Loss curves

REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," (in eng), *CA Cancer J Clin*, vol. 71, no. 3, pp. 209-249, May 2021, doi: 10.3322/caac.21660.

[2] "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395-409, doi: 10.1056/NEJMoa1102873.

[3] S. G. Armato, 3rd *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans," (in eng), *Med Phys*, vol. 38, no. 2, pp. 915-31, Feb 2011, doi: 10.1118/1.3528204.

[4] G. Hinton, A. Krizhevsky, and I. Sutskever, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, pp. 1097-1105, 01/01 2012, doi: 10.1145/3065386.

[5] A. Dosovitskiy *et al.*, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020.

[6] A. Hatamizadeh *et al.*, *UNETR: Transformers for 3D Medical Image Segmentation*. 2021.

[7] G. Szumel *et al.*, *The Impact of Scanner Domain Shift on Deep Learning Performance in Medical Imaging: an Experimental Study*. 2024.

[8] Z. Xue, F. Yang, S. Rajaraman, G. Zamzmi, and S. Antani, "Cross Dataset Analysis of Domain Shift in CXR Lung Region Detection," (in eng), *Diagnostics (Basel)*, vol. 13, no. 6, Mar 11 2023, doi: 10.3390/diagnostics13061068.

[9] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. Khan, and M.-H. Yang, *Intriguing Properties of Vision Transformers*. 2021.

[10] J. Miao, M. Zhang, Y. Chang, and Y. Qiao, "Transformer-Based Recognition Model for Ground-Glass Nodules from the View of Global 3D Asymmetry Feature Representation," *Symmetry*, vol. 15, no. 12, p. 2192, 2023. [Online]. Available: https://www.mdpi.com/2073-8994/15/12/2192.

[11] M. Oumlaz, Y. Oumlaz, A. Oukaira, A. Z. Benelhaouare, and A. Lakhssassi, "Advancing Pulmonary Nodule Detection with ARSGNet: EfficientNet and Transformer Synergy," *Electronics*, vol. 13, no. 22, p. 4369, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/22/4369.

[12] M. A. Mahmoud, Y. Wen, Y. Liufu, X. Pan, R. Su, and Y. Guan, "A Lightweight Dual-Output Vision Transformer for Enhanced Lung Nodule Classification Using CT Images," (in eng), *Technol Cancer Res Treat*, vol. 24, p. 15330338251370439, Jan-Dec 2025, doi: 10.1177/15330338251370439.

[13] K. Cao, H. Tao, and Z. Wang, "Three-Dimensional Multifaceted Attention Encoder–Decoder Networks for Pulmonary Nodule Detection," *Applied Sciences*, vol. 13, no. 19, p. 10822, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/19/10822.

[14] N. Yang, X. Jia, C. Hu, Y. Zhang, and L. Lyu, "A dual-branch encoder context-aware fusion network for ultrasound image segmentation," *Applied Soft Computing*, vol. 182, p. 113538, 2025/10/01/ 2025, doi: https://doi.org/10.1016/j.asoc.2025.113538.

[15] H. Kuang *et al.*, "Hybrid CNN-Transformer Network With Circular Feature Interaction for Acute Ischemic Stroke Lesion Segmentation on Non-Contrast CT Scans," (in eng), *IEEE Trans Med Imaging*, vol. 43, no. 6, pp. 2303-2316, Jun 2024, doi: 10.1109/tmi.2024.3362879.

[16] A. Pal, H. M. Rai, J. Yoo, S. R. Lee, and Y. Park, "ViT-DCNN: Vision Transformer with Deformable CNN Model for Lung and Colon Cancer Detection," (in eng), *Cancers (Basel)*, vol. 17, no. 18, Sep 15 2025, doi: 10.3390/cancers17183005.

[17] H. Jin, C. Yu, J. Zhang, R. Zheng, Y. Fu, and Y. Zhao, "Multitask Swin Transformer for classification and characterization of pulmonary nodules in CT images," (in eng), *Quant Imaging Med Surg*, vol. 15, no. 3, pp. 1845-1861, Mar 3 2025, doi: 10.21037/qims-24-1619.

[18] Y. Wen *et al.*, "Predicting pathological subtypes of pure ground-glass nodules using Swin Transformer deep learning model," (in eng), *Insights Imaging*, vol. 16, no. 1, p. 223, Oct 17 2025, doi: 10.1186/s13244-025-02113-3.

[19] J. Ko, S. Park, and H. G. Woo, "Optimization of vision transformer-based detection of lung diseases from chest X-ray images," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 191, 2024/07/08 2024, doi: 10.1186/s12911-024-02591-3.

[20] H. Mkindu, L. Wu, and Y. Zhao, "Lung nodule detection in chest CT images based on vision transformer network with Bayesian optimization," *Biomedical Signal Processing and Control*, vol. 85, p. 104866, 2023/08/01/ 2023, doi: https://doi.org/10.1016/j.bspc.2023.104866.

[21] Y. Huang *et al.*, "A novel approach to integrating Vision Transformers and machine learning for robust lung nodule classification using CT imaging," *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 3, p. 101672, 2025/09/01/ 2025, doi: https://doi.org/10.1016/j.jrras.2025.101672.

[22] X. Wang *et al.*, "Enhanced pulmonary nodule detection with U-Net, YOLOv8, and swin transformer," (in eng), *BMC Med Imaging*, vol. 25, no. 1, p. 247, Jul 1 2025, doi: 10.1186/s12880-025-01784-0.

[23] K. Sun, Y. Wang, and H. Zhou, "Enhanced pulmonary nodule detection using a transformer framework with dual fusion and gated mechanism," *Complex & Intelligent Systems*, vol. 11, no. 9, p. 397, 2025/07/25 2025, doi: 10.1007/s40747-025-02032-2.

[24] H. Eman, F. Shaukat, M. H. Zafar, and S. M. Anwar, "EMeRALDS: Electronic Medical Record Driven Automated Lung Nodule Detection and Classification in Thoracic CT Images," *Journal of Imaging Informatics in Medicine*, pp. 1-15, 2026.

[25] S. M. Turjya and M. Fawakherji, "Federated lung nodule segmentation using a hybrid transformer–U-Net architecture," *Scientific Reports*, vol. 16, no. 1, p. 5228, 2026/01/14 2026, doi: 10.1038/s41598-026-35243-9.

[26] H. Eman, "Early Detection and Classification of Lung Cancer using Segment Anything Model 2 and Dense Net," *International Journal of Innovations in Science & Technology*, vol. 7, no. 1, pp. 476-492, 2025.

[27] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024/01/22 2024, doi: 10.1038/s41467-024-44824-z.

[28] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015-4026.

[29] M. F. McNitt-Gray *et al.*, "The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation," (in eng), *Acad Radiol*, vol. 14, no. 12, pp. 1464-74, Dec 2007, doi: 10.1016/j.acra.2007.07.021.

[30] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.