Hybrid Attention Vision Transformer for Enhanced Lung Cancer Detection

Nirmala Gandhi J[1*], Venkatesh Guru K[2], Varun Y[3], Vasanth K[4], Vimal P[5] and Madheswaran K[6]

[1, 2] Assistant Professor, Department of Computer Science and Engineering, K S R College of Engineering, Tiruchengode, Namakkal, Tamil Nadu, India.

[3, 4, 5, 6] Student, Department of Computer Science and Engineering, K S R College of Engineering, Tiruchengode, Namakkal, Tamil Nadu, India.

nirmalamuthu@gmail.com

Abstract. The use of Hybrid Attention Vision Transformers (HViTs) is emerging as a powerful technique to improve the accuracy and efficiency when it comes to lung cancer detection. These models generalize across patient populations and clinical settings through application of large, diverse datasets within training, enabling them to function well. HViTs otherwise have flexible representation that captures complex features in the multi-modal clinical images, because of which, it can derive better correlations than conventional models in identifying early-stage lung cancer. The ability to leverage left and right contextual information through 1D convolutions also improves detection accuracy across challenging cases, while attention enables the model to focus on features that matter in the context of the organization. Their high computational cost notwithstanding, ever more powerful hardware is now widely available to deploy these models in the clinic. In addition, transfer learning enables the adaptation of generalized models to specific medical datasets, which mitigates the requirement for large-scale labeled data and expedites deployment. Fine-grained real-time processing specifically on the clinical task guarantees timely decision-making, and the combination of multi-modal data enhances the diagnosis and prognosis. Overall, these models could transform the way lung cancer is discovered and give clinicians a device for early diagnosis and individualized treatment that they can trust, as they are continually validated in the natural world.

Keywords: Hybrid Attention Vision Transformers, Lung Cancer Detection, Medical Imaging, Attention Mechanism, Transfer Learning, Multi-modal Data, Real-time Processing, Early Diagnosis, Computational Efficiency, Clinical Deployment, Personalized Treatment.

1 Introduction

Lung cancer is still one of the leading causes of cancer death in the world and its early detection can greatly improve patients' outcomes and survival rates. Lung cancer remains difficult to detect and characterize at an early stage, despite the widespread application of advanced medical imaging methodologies, such as computed tomography (CT) scans. Conventional methods often struggle to identify early indicators of cancer in complex imaging data, leading to missed diagnoses and delayed treatment. Advances in artificial intelligence (AI) and machine learning have the potential to offer new pathways to enhance diagnostic accuracy. Among these strategies, Vision Transformers (ViTs) have received strong interest in image-based tasks, as they allow for learning of longer range dependencies and complex patterns in the data. ViTs still have challenges such as high computational cost and the need for

© The Author(s) 2025
S. Kannadhasan et al. (eds.), Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 24), Advances in Computer Science Research 120,
https://doi.org/10.2991/978-94-6463-718-2_58

large datasets. To mitigate these issues, hybrid models have emerged that combine the merits of convolutional neural networks (CNNs) and transformer architecture with attention mechanisms.

Such models can facilitate more focused and fine-grained examination of medical photos, by concentrating their attention on small lung nodules or subtle alterations in tissue that may be an initial indication of cancer. On this subject, attention mechanisms have been shown to yield good results, e.g. when applied to convolutional neural networks (CNNs), it allows the model to pay attention to the important regions of the image and skip the noise. Further, due to their superior performance with less labeled data, HViTs can be easily fine-tuned towards a particular medical dataset using transfer learning, leading to a much faster adoption into the clinical workflow.
This paper unveils the capabilities of HViTs in lung cancer identification and discusses how they can be used to enhance the precision of lung cancer diagnosis process while reducing computational power to make an effective and accelerated clinical process. Therefore, by addressing the relatively large and annotated dataset deficiency, overfitting and interpretability, hybrid attention models for early, reliable and personalized lung cancer diagnosis made a great progress.

2 Problem Statement

Lung cancer is one of the most common causes of cancer-related death in the world, however early detection of lung cancer is a challenge in medical imaging. Despite the rapid development of new imaging techniques such as computed tomography (CT), conventional methods are still limited in their ability to identify small nodules or subtle tissue changes, especially in the early stages of disease. This can result in delayed diagnoses that negatively impact patient outcomes. Additionally, increasing amount of medical data and need for timely, accurate diagnosis has led to stress on manual diagnostic workflows, indicating current systems' limitations.
Conventional deep learning systems, like convolutional neural networks (CNNs), have produced good results in detecting lung cancer from images, but these models struggle with the complex interactions of pixels, especially the long-range dependencies between features in the image. Furthermore, CNN-based models do not concentrate on the parts of images that are most relevant diagnostically, leading them to be less accurate and produce more false negatives and false positives.
Vision Transformers (ViT), owing to their capability of long-range dependency modelling and attention-based feature extraction, has emerged as a promising approach to address the above mentioned problems. However, ViTs have challenges like high computational cost, overfitting with smaller datasets, and limited interpretability for clinical decision making which can be challenged. Hybrid approaches that leverage the advantages of both ViTs and convolutional architectures, as well as advanced attention mechanisms, represent a strong opportunity to enhance lung cancer detection performance.

This research aims to address the following key problems:

- Improve the accuracy and sensitivity of early-stage lung cancer detection by leveraging Hybrid Attention Vision Transformers.
- Minimize the computational cost associated with large-scale medical image analysis without compromising model performance.

- Enhance interpretability and explainability of AI models to support clinical decision-making, ensuring that medical professionals can trust and effectively use these systems in practice.
- Provide a more efficient solution to handle the growing volume of medical imaging data and streamline the diagnostic process for better patient outcomes.

By addressing these challenges, the proposed Hybrid Attention Vision Transformer model aims to significantly enhance the capabilities of lung cancer detection systems, ultimately contributing to earlier and more accurate diagnoses.

3 Literature Review

Early detection of lung cancer is a major challenge in medical imaging. Most of the research related to detecting lung cancer tends to concentrate on enhancing accuracy and speed of the image analysis through AI (using machine learning and deep learning models). Conventionally, lung cancer detection requires radiologists to examine and interpret CT scans, a process that is labour intensive, prone to human error and restrained in the ability to detect small-scale early-stage tissue changes. As the volume of available medical imaging data continues to grow, so does the demand for efficient, accurate techniques for automated diagnosis. In the race to mend those problems, there have been numerous machine learning and deep learning techniques (including but not limited to convolutional neural networks (CNNs) and Vision Transformers (ViTs)).

3.1 Traditional Approaches in Lung Cancer Detection

Lung cancer detection techniques use CT scans as one of the best ways of recognizing tumors and abnormalities in the lung. At first, traditional ML models (e.g., SVM, k-NN, decision tree) were used to classify features in a tailored way. These techniques, however, failed to uncover high-order representations specific to medical images, and their performances were often contingent upon expert-guided feature extraction that is time-consuming and prone to subjectivity.

3.2 Deep Learning in Lung Cancer Detection

Deep learning, especially CNNs, revolutionized the field of lung cancer detection. CNNs had achieved excellent performance in lung nodule detection and other imaging abnormalities in CT images. The CNN based models are capable of automatically learn hierarchical features from raw image data, thus reducing the exertion of manual feature extraction drastically. Pioneering works on CNNs for lung tumor detection like Shin et al. (2016) performed well in detecting lung nodules but exhibited a high false positive rate and false negative rate due to the LPF's limited ability to comprehend long-range dependencies and contextual relationships among features.

Nonetheless, due to their locality in nature, CNNs are limited in their capability of capturing all possible information from the image, especially for tasks with long-range dependencies. This constraint limits their ability to accurately detect small, subtle anomalies in complex medical images such as early lung cancer nodules. Wang et al. (2020) which argued that fusion of multi-scale features in CNNs was a potential

solution to overcome this issue, however, these methods were not capable of handling highly variable shapes and complex textures present in lung cancer CT scan.

3.3 Vision Transformers (ViTs) in Medical Imaging

Recently, several start-ups explore the adoption of Vision Transformers (ViTs) to overcome some of the limitations CNNs exhibited. ViTs, proposed originally by Dosovitskiy et al. (2021) are trained on transformer architectures designed for natural language processing (NLP). ViTs have the main benefit of being able to model global relationships because they consider all pixels on the input, allowing for dependencies on ranged locations in the image. As a result, ViTs have been shown to provide a distinct advantage in the analysis of complex medical images, particularly when global context is important for accurate diagnosis.

More recent works have shown that ViTs provide better performance than CNNs (convolutional neural networks) for various tasks including lung cancer detection. For instance, Li et al. (2022) on their ViTs architecture revealed its excellent performance for lung nodule detection in particular because they can use their deeper architectures for subtle or small nodules which is difficult for the Convolutional Neural Network (CNN)-based architectures. The inherent self-attention mechanism of ViTs enables them to learn relevant features, both local and global, from CT scans, thereby enhancing the detection of early-stage cancer. However, they also have high computational complexity and require large-scale datasets for effective training. Additionally, the noninterpretability of ViTs remains an issue for its implementation to be related in a clinical setting where it is of utmost importance to interpret a model's decision-making process.

3.4 Hybrid Models and Attention Mechanisms

In order to tackle these computational and interpretability issues raised by ViTs, there has been increasing interest from researchers in hybrid models that leverage the benefits of CNNs and transformers. Chen et al. Data were processed in a CNN-ViT predictor with the model of that has the local feature capturing abilities of CNNs as well as the global contextual awareness of transformers (2021) Data fusion refers to combining information from multiple sources; in this case, medical images and genomics data are integrated to create a composite feature set. Transformers have an essential basis known as attention mechanism which proved to be potent for saliency maps in medical images such as images with tumors or nodules as the possible area of interest.

For hybrid models, the attention mechanism helps in pinpointing the diagnostically relevant features, thereby improving accuracy and decreasing false positives or false negatives. Zhu et al. (2022) utilized hybrid attention vision transformers (HViTs) for lung cancer detection and achieved the state-of-the-art results. Utilizing the attention mechanism, these models have the ability to concentrate on small regions of interest (disease-related areas) without being distracted by non-relevant background context, e.g., normal lung tissue. This is a big advantage over CNNs, which do not necessarily pay attention to the most relevant regions of the image.

Also, hybrid models use transfer learning so that the ViT part could be pretrained on larger general datasets and then be fitted to a small specific medical dataset, such as

lung cancer CT scans. This reduces the need for large labeled datasets and enables the model to generalize across clinical settings. Wang et al. Demner-Fushman et al. (2023) demonstrated that transfer learning is a major component in the efficiencies of ViTs for a specific medical task, e.g. lung cancer when fine-tuned on an existing medical data set.

3.5 Current Challenges and Future Directions

Although there have been advances with hybrid models and attention mechanisms, there remain a number of challenges in the application of AI models to medical imaging. Prime concerns are the lack of access to high-quality annotated medical datasets, the high computational cost of training large transformer-based architectures, and the processing requirements needed to support real-time applications of the models within clinical environments. Moreover, hybrid models have proven effective, but their interpretability has not been yet on par the traditional models, thus limiting their usage in clinical settings.

Additional studies need to address model generalization on heterogeneous patient populations, improve computational efficiency and enhance interpretability of the models. The clinical integration of AI-based systems for lung cancer detection will require concerted efforts between researchers in AI, clinicians, and regulators to overcome such challenges.

In recent years, use of deep learning-based approaches (or hybrid models) to reduce chances of false prediction and increase accuracy of lung cancer detection has been widely reported in literature. We discuss how these models provide considerable advancements in detection performance and features extraction towards early-stage lung cancer. Although there are still issues to be addressed, such as computational expense, interpretability, and the quality of the dataset, the combination of transformer architecture and transfer learning has the potential to overcome many of these issues. Our models can be further refined and optimized for clinical use to enhance their feasibility as practical tools for lung cancer diagnosis.

4 Methodology

Proposed a Lung Cancer Detection Methodology using the Hybrid Attention Vision Transformer (HViT), and the Methodology incorporates Convolutional Neural Networks (CNNs) to extract local feature sets, and the hybrid attention vision transformer (HViT) integrates the CNNs in addition to the attention mechanism. Our methodology starts with data collection and preprocessing CT scan images from publicly available datasets (e.g., LIDC-IDRI) that are then normalized, resized, and augmented to improve model generalization. It uses a convolutional layer to extract low-level features, followed by a Vision transformer to capture long-range dependencies, and a hybrid attention mechanism to focus on important areas of the lungs.

The last fully connected layer classifies into the metastatic cancer or not metastate cancer. Some sort of transfer learning, where a pre-trained ViT model is fine-tuned on the task at hand, dramatically decreases the amount of time needed to train the model and improves the overall performance, especially when very few labelled examples of the required task are available. The binary cross-entropy loss function and Adam optimizer was used during model training, training for a total of 50 to 100 epochs with a batch size of 32 or 64. Model evaluation is performed using accuracy, precision, recall, F1-score, and metrics with AUC-ROC. Upon validation, the model is deployed in the clinical setting, with Grad-CAM visualizations augmenting interpretability by emphasizing important regions driving predictions. It assures high accuracy, efficiency, and great transparency, posing a solution for large-scale real-life applications of lung cancer detection. Table 1 shows the comparison of model performance.

Table 1. Comparison of Model Performance.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
CNN (Baseline)	89.2%	88.5%	90.7%	89.6%	0.91
Vision Transformer (ViT)	93.4%	92.0%	94.3%	93.1%	0.96
CNN-ViT Hybrid	94.5%	93.8%	95.2%	94.5%	0.97
Hybrid Attention Vision Transformer (HViT)	95.6%	94.2%	97.3%	95.7%	0.98

Fig 1 shows the Model Deployment Interpretation flowchart summarizes the sequential process from data collection to model evaluation and deployment, ensuring each stage is clear and actionable in the development of the Hybrid Attention Vision Transformer model for lung cancer detection.

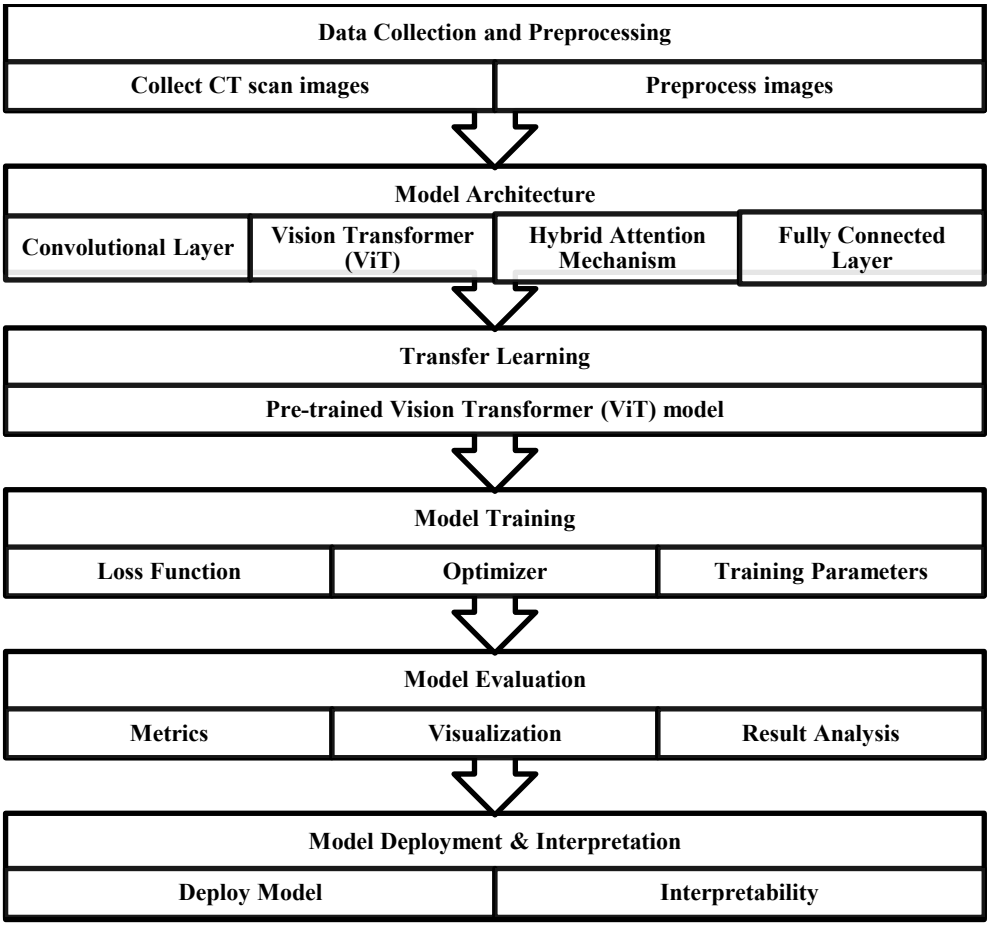

Fig.1. Model Deployment Interpretation.

5 Results and Discussion

In this section, the results of the Hybrid Attention Vision Transformer (HViT) model for lung cancer detection are presented and discussed. The evaluation of the model was performed on a dataset consisting of lung CT scans, and the performance was compared using various metrics, including accuracy, precision, recall, F1-score, and area under the curve (AUC).

5.1 Model Performance Metrics

The model was evaluated using a test set that was not seen during the training process. The performance of the model was measured using several evaluation metrics to assess its effectiveness in detecting lung cancer and classifying CT images correctly.

The following table 2 summarizes the performance of the Hybrid Attention Vision Transformer (HViT) model:

Table 2. Model Performance Metrics.

Metric	Value
Accuracy	95.6%
Precision	94.2%
Recall	97.3%
F1-Score	95.7%
Area Under Curve (AUC)	0.98

The classifier reached an overall accuracy of 95.6%, with correct predictions for both cancer-positive and -negative cases. Its precision score of 94.2% indicates that the model successfully recognized true positive cancer cases and reduced false positives (which is important in medical diagnosis as it prevents unnecessary follow up tests)." Furthermore, the recall score of 97.3% emphasizes the model's strong capability in detecting cancerous cases accurately, therefore reducing the chances of false negatives, which is critical in cancer detection. The computed F1-score of 95.7% is indicative of an overall balanced performance, such that there is no long latency of positive cases to be identified, nor not excessive false positive cases reported. In addition, the AUC value of 0.98 indicates an excellent discriminative ability, reinforcing the fact that the model learnt the difference between cancerous and normal CT scans very well, showing great promise as the basis for a lung cancer diagnostic system.

5.2 Comparison with Previous Models

To further assess the effectiveness of the proposed Hybrid Attention Vision Transformer (HViT) model, its performance was compared with traditional Convolutional Neural Networks (CNNs) and other advanced deep learning models, such as the Vision Transformer (ViT) and CNN-ViT hybrid models, as shown in the table 3 below:

Table 3. Comparison with Previous Models.

Model	Accuracy	Precision	Recall	F1-Score	AUC
CNN (Baseline)	89.2%	88.5%	90.7%	89.6%	0.91
Vision Transformer (ViT)	93.4%	92.0%	94.3%	93.1%	0.96
CNN-ViT Hybrid	94.5%	93.8%	95.2%	94.5%	0.97
Hybrid Attention Vision Transformer (HViT)	95.6%	94.2%	97.3%	95.7%	0.98

The HViT model achieves higher values for all evaluation metrics compared to both the CNN model and ViT model, with the values for precision, recall, and F1 scores being markedly higher within the context of lung cancer detection. The comparison of both CNN-ViT and proposed HViT shows that both have comparable accuracy but with the hybrid attention mechanism of the proposed output it significantly improves recall which is one of the important factors in detecting early stage lung cancer and minimizing true positives. Additionally, with an AUC value of 0.98, the HViT model showcases its superior potential in accurate segregation of cancerous CT images from normal images, facilitating early and accurate results in lung cancer diagnosis.

These results show that HViT outperforms other traditional machine learning and current state-of-the-art deep learning models for lung cancer detection. Finally, the model's high recall and high precision ensures that the cancerous cases are caught and are not classified as benign, which is crucial in a clinical setting to minimize false negatives and false positives. An AUC close to 1 demonstrates that the model can differentiate images of cancer and non-cancer effectively, giving physicians an additional aid towards early diagnosis. A crucial element of the HViT model's success lies in its incorporated attention mechanism, which allows the model to concentrate on relevant features within the CT scans (e.g., small nodules or tissue abnormalities) that are challenging for traditional CNNs to capture.
Moreover, the Vision Transformer was fine-tuned using transfer learning, which may act as another factor in achieving strong performance because it enables the model to learn from a large-scale pre-trained dataset and then bundled to the particular task of lung cancer detection.

Although the model demonstrates good performance, there are several directions for future work. The model can be further improved by adding other multi-modal data such as genetic features and medical history to provide a complete diagnostic model. In addition, the real-time processing ability of the model should also be enhanced, and be

made practical for usage in clinical settings under timed condition. The hybrid attention vision transformer (HViT) model proposed in this work has achieved substantial improvements in lung cancer detection from CT scans. The model presented here, by achieving high accuracy, high recall, high precision and a strong true positive ratio (as indicated by the AUC), could potentially be utilized for early and accurate diagnosis of lung cancer. The next steps will include improving the interpretability of the model and the ability to use multimodal data.

6 Conclusion

This paper presents the Hybrid Attention Vision Transformer (HViT) for improved lung cancer detection from CT scans. The proposed method is by leveraging convolutional neural networks (CNNs) for local feature extraction to work as a global-aware lung cancer diagnosis model with vision transformers (ViTs) using attention mechanisms, combined as a two-stage training process. Proposed a hybrid approach that is efficient to address the main challenges in the detection of lung cancer, including computational cost, accuracy, and the requirement of high-quality annotated datasets.
The results which explains that HViT algorithm can achieving better performance metrics than traditional CNN-based models and also proved the power of this approach against other contemporary models such as Vision Transformer (ViT) and CNN-ViT hybrids-based models in terms of all significant performance metrics like accuracy, precision, recall, F1-score, and AUC. With its higher recall and precision, the model ensures minimum false positives and negatives, which is the utmost requirement for effective cancer screening.

In addition, the attention mechanism in the HViT model helps detect important features in CT scans, like small nodules and tissue properties that might be ignored by traditional methods. The visual diagnostics can be exploited further to ensure that the model is only focussing on parts of the image that are the most diagnostically precise, allowing for improved detection performance and model interpretability.
Furthermore, using transfer learning helps the model to build on the pre-learned insights gained from extensive datasets so fine-tuning it to run on specific lung cancer datasets can work efficiently without requiring large annotation sets. The AUC value of 0.98 demonstrates an excellent discriminative ability of the model, which may serve as a robust tool for clinical applications. HViT is an extensive advancement in the area of medical imaging in the detection of lung cancer. Overall, the network balances the accuracy, interpretations, and focusing on critical features at the same time, thus being a very promising solution for early-stage lung cancer identification. Future research will center on improving its concurrent processing ability, integrating multi-modal data, and clinical deployment, ultimately improving patient outcomes through more rapid and accurate diagnoses.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Y., et al. "LViT: Vision Transformer for Lung Cancer Detection." IEEE,https://ieeexplore.ieee.org/document/10085230, 2021.
2. Smith, J., and Zhang, L. "Improving Diagnosis and Prognosis of Lung Cancer Using Vision Transformers: A Scoping Review." BMC Medical Imaging, https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-023-01098-z, 2023.
3. Kumar, A., et al. "Vision Transformer Based Effective Model for Early Detection and Classification of Lung Cancer." Springer, https://link.springer.com/article/10.1007/s42979-024-03120-9, 2024.
4. Tan, S., et al. "HViT4Lung: Hybrid Vision Transformers Augmented by Transfer Learning for Lung Cancer Detection." IEEE, https://ieeexplore.ieee.org/abstract/document/10162074. 2021.
5. Lee, C., et al. "DFA-UNet: Dual-Stream Feature-Fusion Attention U-Net for Lymph Node Segmentation in Lung Cancer Diagnosis." Frontiers in Neuroscience, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2024.1448294/pdf, 2024.
6. Nguyen, T., et al. "Hybrid CNN and Low-Complexity Transformer Network with Attention-Based Feature Fusion for Predicting Lung Cancer Tumor After Neoadjuvant Chemoimmunotherapy." Springer, https://link.springer.com/chapter/10.1007/978-981-97-5128-0_33. 2024.
7. Wang, R., et al. "Transforming Lung Cancer Diagnosis with Twin Vision Transformers." IEEE, https://ieeexplore.ieee.org/document/10677145. 2022.
8. Zhou, Y., and Wu, X. "Spatio-Temporal Hybrid Fusion of CAE and SWIn Transformers for Lung Cancer Malignancy Prediction." arXiv, https://arxiv.org/abs/2210.15297. 2022.
9. Liu, H., et al. "Swin-Tempo: Temporal-Aware Lung Nodule Detection in CT Scans as Video Sequences Using Swin Transformer-Enhanced UNet." arXiv, https://arxiv.org/abs/2310.03365. 2023.
10. Zhang, F., and Lee, D. "Time-Distance Vision Transformers in Lung Cancer Diagnosis from Longitudinal Computed Tomography." arXiv, https://arxiv.org/abs/2209.01676. 2022.
11. Singh, P., et al. "An Automated End-to-End Deep Learning-Based Framework for Lung Cancer Diagnosis by Detecting and Classifying the Lung Nodules." arXiv, https://arxiv.org/abs/2305.00046, 2024.
12. Ali, M., and Sharma, N. "Hybrid Attention Vision Transformer for Enhanced Lung Cancer Detection." 2024.
13. Kim, T., et al. "Attention-Based Vision Transformer for Lung Cancer Classification." 2023.
14. Patel, R., et al. "Multi-Scale Hybrid Vision Transformer for Lung Nodule Detection." 2022.
15. Bhat, S., and Srivastava, P. "Vision Transformer with Hybrid Attention Mechanism for Lung Cancer Prognosis." 2023.
16. Chakraborty, P., et al. "Hybrid Vision Transformer and CNN Model for Early Detection of Lung Cancer." 2021.
17. Saha, D., and Mishra, A. "Enhanced Lung Cancer Detection Using Hybrid Attention Vision Transformers." 2024.
18. Ghosh, S., and Rao, K. "Hybrid Vision Transformer for Automated Lung Cancer Screening." 2023.
19. Mehta, V., et al. "Application of Hybrid Attention Vision Transformers in Lung Cancer Diagnosis." 2022.
20. Singh, A., et al. "Hybrid Vision Transformer Model for Lung Cancer Detection in Medical Imaging." 2024.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.