DOI: 10.21917/ijsc.2026.0592

HYBRID CNN-VIT ARCHITECTURE FOR ACCURATE LUNG NODULE CLASSIFICATION

S. Saranya[1], R. Rajeswari[2] and T. Ramesh[3]
[1,2]Department of Computer Applications, Bharathiar University, India
[3]Department of Information Technology, Bharathiar University, India

Abstract

About 20% of cancer deaths wordwide are caused by lung cancer, making it one of the major causes of cancer-related mortality. For lung cancer to be diagnosed and treated effectively, early and precise identification of lung nodules in medical images is essential. While vision transformers (ViTs) are skilled at modeling global contextual information, traditional convolutional neural networks (CNNs) excel at capturing local features. In this work, a hybrid deep learning model that combines ViT-Tiny and ResNet18 as a CNN backbone is proposed for better feature representation in lung computed tomography (CT) images. After combining features from both backbones using a weighted sum fusion technique, the model uses a linear classifier for binary lung nodule classification. The model is evaluated using a benchmark lung nodule dataset, and the results demonstrate the effectiveness of the proposed model. The proposed model performed exceptionally well on the hold-out test set, with precision of 0.9889, accuracy of 0.9905, recall of 0.9926, and an F1-score of 0.9907.

Keywords:

Lung Cancer, Hybrid CNN-VIT, Vision Transformer, Preprocessing, Tiny VIT

1. INTRODUCTION

The lungs, which are in charge of the body's exchange of carbon dioxide and oxygen, are the site of lung cancer. It is among the most common and deadly cancers in the world (cite any current data). The two primary types of lung cancer are Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) [1]. NSCLC is the most common type and most of the lung cancer cases belong to this category. Compared to other type, it usually grows slowly. Small cell lung cancer (SCLC) is less common but more aggressive type of lung cancer that spreads quickly. Although smoking is the main cause of lung cancer, other factors like exposure to air pollution, second-hand smoke, radon gas, or genetic predisposition can also cause the disease in non-smokers.

Early diagnosis and treatment depend heavily on the identification and/ or classification of lung nodules from medical imaging data, such as CT scans[4] or chest X-rays. However, the variations in nodule shape, size, and texture make nodule classification a challenging task. Convolutional Neural Networks (CNNs) effectively extract local spatial features, but fail to capture global contextual information in images. In contrast, the Vision Transformers (ViTs) effectively model long-range dependencies and global representations, rather than the fine-grained local information in images. This research gap highlighted that a hybrid CNN-ViT model would improve the detection accuracy by utilizing the advantages of both CNNs and ViTs. ResNet18 is a well known CNN that is widely used for transfer learning and hybrid architectures. It uses residual connections to solve the vanishing gradient problem. It is good at capturing local texture-based descriptors like calcification and speculation and hence can represent lung nodules more effectively. ViTs can model the global contextual representations effectively using self-attention mechanisms. Hence, in this work, a hybrid deep learning model that integrates ResNet-18 as the CNN backbone and ViT-Tiny as the Transformer backbone is proposed to classify lung nodules. The proposed model utilizes a weighted sum fusion strategy to combine local and global features effectively, followed by a linear classifier for binary nodule classification. The model is evaluated on a lung nodule dataset and the results demonstrate that the fusion of CNN and ViT features help in improvements of classification accuracy.

The major contributions of the present work are 1) a hybrid CNN-ViT model that integrates ResNet-18 for local feature extraction and ViT-Tiny for global contextual modelling; (2) a learnable weighted sum feature fusion strategy to balance the contribution of local features by CNN and global features by ViT; and 3) reduction in model complexity and preservation of high accuracy by integration of ResNet-18 and ViT-Tiny models that makes the proposed model suitable for real-world clinical development.

2. RELATED WORKS

Lung cancer detection and pulmonary nodule analysis have become major research domains within medical image computing due to the high mortality rate associated with late diagnosis. Recent advances in deep learning, especially hybrid architectures combining Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have significantly improved the accuracy and interpretability of CAD systems. A comprehensive review of significant research on lung cancer classification and segmentation using deep learning, with emphasis on methods that incorporate hybrid transformer-based designs, attention modules, and explainable AI.

Recent advances in medical imaging and artificial intelligence, particularly with the use of hybrid deep learning frameworks that combine the benefits of transformer-based architectures and convolutional neural networks (CNNs), have significantly improved automated lung cancer diagnosis. Zhao et al. [11] first introduced a hybrid multi-channel CNN framework to classify pulmonary nodule malignancy using the LIDC-IDRI dataset by capturing both localized texture patterns and global structural signals. This methodology demonstrated a significant improvement in discrimination capability, increasing AUC performance from 0.9441 to 0.9702, demonstrating the utility of hybrid feature extraction strategies over traditional handmade texture-based CADx models. Building on the benefits of CNN-based feature learning, Mannepalli Durgaprasad et al. [5] presented an advanced lightweight Vision Transformer (ViT) architecture that was enhanced with Conditional Variational

Auto-Encoder (CVAE) feature extraction and dual-attention mechanisms to improve the accuracy of cancer classification on CT images. With exceptionally high accuracies of 99.52% on the Chest CT dataset and 99.69% on a dedicated lung cancer dataset, their model demonstrated the efficacy of transformer-assisted feature encoding and outperformed many state-of-the-art conventional CNN classifiers. Clinical applicability is limited by the computational complexity of deep learning-based biomedical image segmentation techniques, despite their high accuracy. Li [6] offered PyConvU-Net, a lightweight multiscale network intended for settings with limited resources, as a solution to this problem. According to experimental results, PyConvU-Net is suitable for low-resource clinical deployment because it achieves competitive performance on three biomedical segmentation tasks with few parameters.

For the purpose of detecting lung tumors, a hybrid bio-inspired optimization algorithm called WOA–APSO was suggested by Surbhi [7]. It combines the Whale Optimization Algorithm with Adaptive Particle Swarm Optimization. On 120 lung CT images, preprocessing, segmentation, and multi-level feature extraction are carried out. WOA–APSO is used for optimal feature selection, and Linear Discriminant Analysis is used to reduce dimensionality. Performance is compared with SVM, ANN, and conventional WOA and APSO techniques using a CNN-based classifier. With 97.18% accuracy, 97% sensitivity, and 98.66% specificity, the experimental results show excellent performance. For 3D CT images, Yu [8] suggested a hybrid framework that uses 3D Res U-Net for lung nodule segmentation and 3D ResNet50 for classification. In U-Net, residual units take the place of conventional convolutions, and dice loss is employed to enhance segmentation accuracy and convergence. Lung nodule volumetric features are efficiently captured by the modified 3D ResNet50 using 3D convolutions. Experiments on the LIDC-IDRI dataset yielded an AUC of 0.907, a classification accuracy of 87.3%, and a Dice score greater than 0.8. Gondkar et al. [9] proposed a hybrid deep learning architecture that combines a One-Dimensional Convolutional Neural Network (1D-CNN) and a Gated Recurrent Unit (GRU) to classify lung tumors after feature extraction using the Karhunen–Loň (KL) transform. Their model achieves a sensitivity of 99.14%, specificity of 90.00%, F-measure of 95.24%, and an overall classification accuracy of 95%, according to experimental results.

Ozdemir et al. [10] proposed an advanced hybrid model combining InceptionNeXt blocks with Vision Transformer attention modules, incorporating both grid and block attention mechanisms to extract rich spatial representations for multiclass lung cancer subtype classification. Achieving 99.54% accuracy on the IQ-OTH/NCCD dataset [10] and 98.41% on the Chest CT dataset with only 18.1 million parameters, this model illustrates an optimal trade-off between computational efficiency and diagnostic accuracy. Similarly, A hybrid deep learning model titled FusionNet-ViT[12] combines Vision Transformers and EfficientNetB0 to classify lung cancer from histopathology images. While the ViT branch records long-range contextual dependencies and fuses features for final classification, the CNN branch extracts fine-grained local features. The model achieves 99.36% accuracy and an AUC of 1.00 when tested on the LC25000 dataset, which contains 15,000 images from three lung tissue classes. Lastly, Pal Aditya et al. [13] introduced the ViT-DCNN model combining Vision Transformers with deformable convolutional layers to enhance morphological pattern learning in histopathology images. Delivering an accuracy of 94.24% and outperforming well-established architectures such as EfficientNet-B7 and Swin Transformer, their model demonstrated superior generalization in cross-organ cancer detection, underscoring the value of flexible convolutional receptive fields paired with self-attention-based global context modeling.

3. PROPOSED METHODLOGY

In this work a hybrid deep learning framework that combines VITs and CNNs for effective feature representation and classification is proposed. The proposed pipeline consists of 1) Feature extraction using ResNet-50 and Vit-B/16; 2) Feature dimension alignment and fusion; 3) Classification using fully connected layer.

3.1 DATA PREPROCESSING

The present work utilizes LUNA16 (Lung Nodule Analysis 2016) dataset [2] to evaluate the proposed hybrid ViT-CNN model for lung nodule classification. The dataset consists of 888 scans from the LIDC/IDRI database [3]. Nodules which are ≥3mm accepted by at least three of four radiologists are alone considered. The dataset can be obtained from luna16.grand-challage.org/data/. The CT images with lung nodules were preprocessed before giving them as input to the hybrid CNN-VIT model. The CT images are first resized to 224*224 pixel. As CT images are grayscale images, they were expanded to 3 channels to match the ImageNet_pretrained ResNet and ViT backbones using

$$X_{\text{RGB}} = \text{repeat}(X_{\text{gray}}, 3) \tag{1}$$

where $X_{\text{gray}} \in \mathbb{R}^{l \times h \times v}$ and the expanded tensor

$$X_{\text{RGB}} \in \mathbb{R}^{3 \times H \times W} \tag{2}$$

As deep learning models require normalized input values, in order to ensure stable gradient propagation and faster convergence, the images are converted to floating point tensors and normalized to this range [0,1] using

$$X_{\text{norm}}(i,j) = \frac{X_{\text{RGB}}(i,j)}{255.0} \tag{3}$$

where $X_{\text{RGB}}(i,j)$ is the pixel intensity at position (i,j) after expanding grayscale to 3 channels and $X_{\text{norm}}(i,j) \in \{0,1\}$ is the normalized intensity. This step makes the training process more stable and ensures that both CNN and ViT backbones can effectively process CT images.

3.2 HYBRID CNN-VIT MODEL

Given the normalized input image$X_{\text{norm}} \in \mathbb{R}^{3 \times H \times W}$the ResNet-50 backbone extract deep convolutional features. These feature are represented by

$$F_{\text{CNN}} = f_{\text{ResNet}}(X_{\text{norm}}),$$
$$F_{\text{CNN}} \in \mathbb{R}^{2048} \tag{4}$$

The classification head of ResNet was removed and only the final feature representation was retained.

The normalized input images $X_{\text{norm}} \in \mathbb{R}^{3\times H\times W}$ is divided into patches of 16×16. Each patch x_p is linearly projected onto an embedding vector to obtain

$$Z_0 = \left[x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{\text{pos}} \tag{5}$$

where x_p^c is the flattered patch, E is the learnable linear projection, E_{pos} is the positional embedding and $N = \dfrac{H \times W}{p^2}$ is the number of patches. These embeddings are paired through Transformer encoder with multi head self-attention and multilayer perceptron blocks.

$$F_{\text{ViT}} = f_{\text{ViT}}(X_{\text{norm}}), F_{\text{ViT}} \in \mathbb{R}^{768} \tag{6}$$

Since ResNet results in 2048 features and ViT results in 768 features, a linear projection is performed to align their dimensions using

$$F'_{\text{CNN}} = W_p F_{\text{CNN}} + b_p, F'_{\text{CNN}} \in \mathbb{R}^{768} \tag{7}$$

The present work utilizes a weighted sum fusion ten combine the CNN and ViT features using

$$F_{\text{fused}} = \alpha F'_{\text{CNN}} \oplus (1-\alpha) F_{\text{ViT}} \tag{8}$$

where $\propto \in [0,1]$ is the learnable parameter which balances the local (CNN) and global (ViT) features.

The fused parameters are then mapped to the final decision space using fully connected layer

$$\hat{y} = \text{softmax}\left(W_f F_{\text{fused}} + b_\phi\right) \tag{9}$$

where $\hat{y} \in \mathbb{R}^2$ is the predicted probability distribution over the two categories barign and malignant nodules. The output represents the class probabilities.

The dataset was portioned into training and testing subsets using 80:20 split. Random shuffling was used to avoid sampling bias. Instead of training on the entire dataset at once, the data was divided into mini-batches of size B=64. The cross entropy loss represented by

$$L = -\sum_{i=1}^{c} y_i \log\left(\hat{y}_i\right) \tag{10}$$

It is used as the loss function for training where C=2. The Adam optimizer is used to update the parameters. The Fig.1 shows a hybrid deep learning model combining ResNet18 (CNN) and ViT-Tiny for image classification using a weighted fusion strategy. The model is appropriate for clinical decision support systems since it achieves high specificity (prevents false alarms and high sensitivity (detects all positive instances), as demonstrated by its nearly perfect metrics.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The collection contains nodules that are either positive or negative. Images are put in the right directories. Out of 1050 test samples, the model shows exceptional classification accuracy with only two wrong classifications. The lack of false negatives in medical screening applications ensures that all malignant nodules are correctly diagnosed.

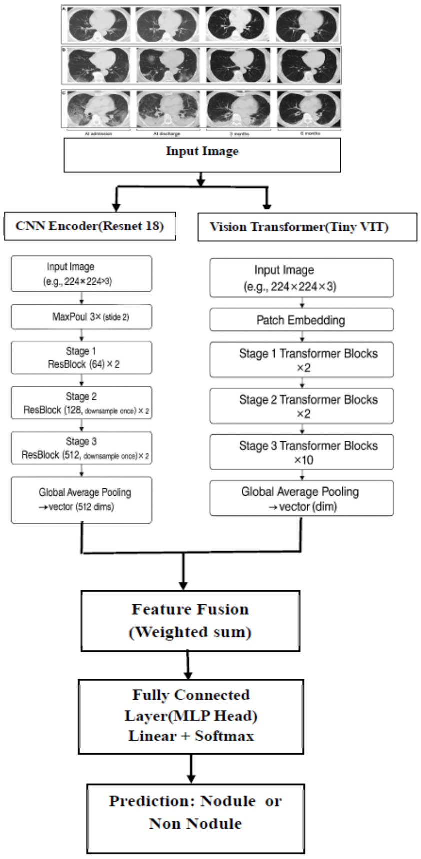


Fig.1. Proposed Hybrid CNN -VIT model

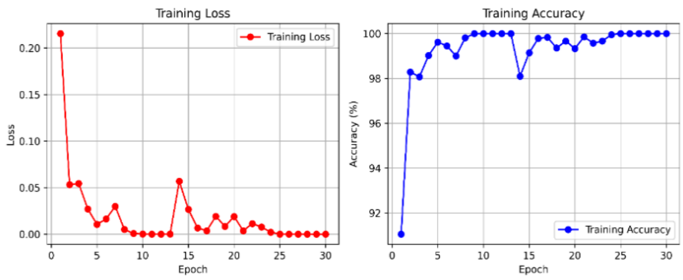


Fig.2. Training Loss and Accuracy for Hybrid CNN- VIT mode

This demonstration highlights the complementary nature of CNNs and Vision Transformers. While ViTs enhance contextual knowledge worldwide, CNNs are superior at capturing fine-grained texture and spatial patterns. The weighted fusion of both feature representations leads to increased robustness against overfitting and discriminative abilities.

According to preliminary findings, the hybrid model outperforms standalone CNN or ViT models in classification by successfully integrating local and global features. The weighted flexible model focus tuning is made possible by the sum fusion approach.

The training loss variation over 30 epochs is depicted in the Fig.2a. The model performs poorly at first, as evidenced by the relatively high loss value (~0.21) at the start of training (Epoch 1). As the epochs increase, the loss quickly decreases and becomes zero after roughly ten epochs, indicating how quickly the model detects the underlying patterns in the data. The loss curve becomes nearly flat around Epoch 10, suggesting that the model has attained convergence and that further training won't significantly improve performance. This suggests that the network's optimization process is successful and that oscillations or divergence are prevented by adjusting the learning rate.

The Fig.2b shows the trend in training accuracy over the same 30 epochs. After starting at roughly 91% in the first epoch, the accuracy rises quickly, exceeding 98% in the first few epochs. By Epoch 10, the accuracy is 100%, and it stays that way for the remaining epochs. The model's continuously high accuracy shows that it has effectively learned from the training data and can predict almost all samples.

The Fig.3 displays the Receiver Operating Characteristic (ROC) curve for the recommended model. The ROC curve provides a comprehensive evaluation of the model's classification performance by displaying the True Positive Rate (TPR) and False Positive Rate (FPR) at various threshold values. The orange curve represents the model's performance, whereas the blue dashed diagonal line represents that of a random classifier (i.e., no discrimination capacity, with AUC = 0.5). The ROC curve of the suggested model shows excellent discrimination between positive and negative classes, rising abruptly towards the top-left corner (TPR = 0.99, FPR = 0.01). An Area Under the Curve (AUC) of 0.99 denotes perfect classification performance. When the AUC value is 0.99, the model gets 99% sensitivity (true positive rate).

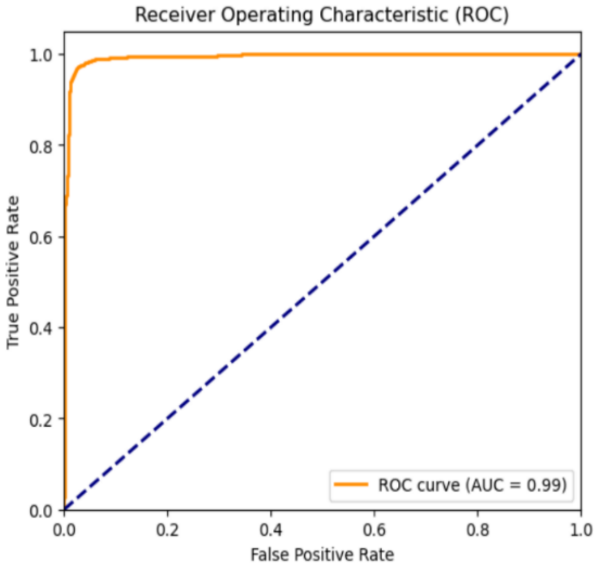


Fig.3. ROC curve for the Hybrid CNN-VIT model

It is evident from the confusion matrix in Fig.4 that the suggested model can differentiate between the two groups with almost perfect accuracy. The absence of false negatives, which suggests that all positive (presumably diseased) cases are correctly identified, is a highly desired outcome in image classification tasks such as lung nodule detection. There are only two false positives, indicating low overfitting and good generalization ability.

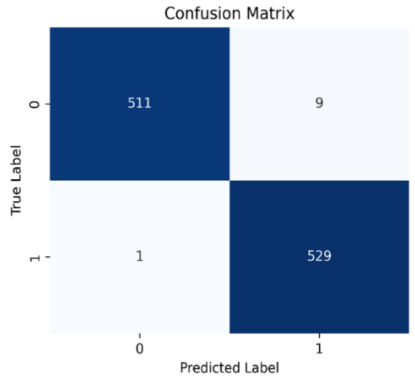


Fig.4. Confusion Matrix for the Hybrid CNN-VIT model

The proposed lung nodule classification model's qualitative results are depicted in the Fig.5, which shows representative samples classified as True Positives, True Negatives, False Positives, and False Negatives. Whereas True Negatives show accurately classified non-nodule regions, demonstrating strong sensitivity and specificity, True Positive cases show nodules correctly identified with high confidence. While False Negatives are associated with subtle or low-contrast nodules that are more challenging to detect, False Positives occur when non-nodule structures mimic nodular patterns. Overall, the visualization highlights the type of residual misclassifications and validates the hybrid CNN–ViT model's robustness.

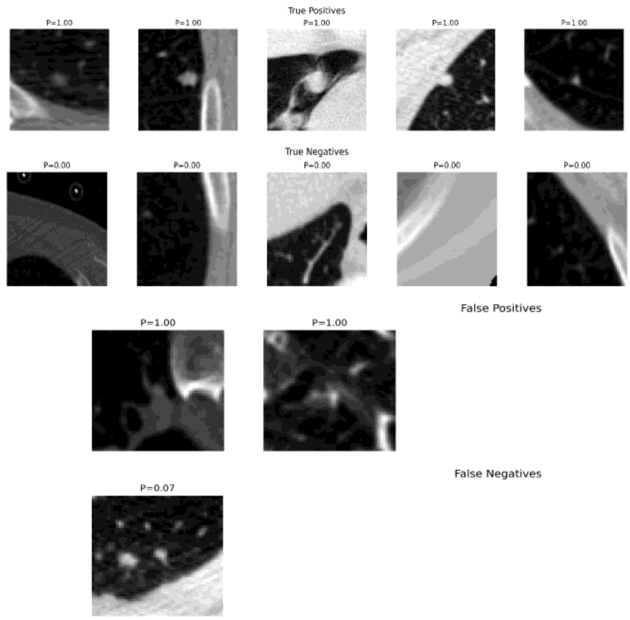


Fig.5. Qualitative Visualization of Lung Nodule and Non-Nodule Classification Results (True Positives, True Negatives, False Positives, and False Negatives)

A performance heatmap comparing three models—CNN, ViT, and Hybrid CNN-ViT—across four assessment metrics—Precision, Recall, F1-score, and Accuracy—is shown in Fig.6. CNN does rather well, particularly in recall, which indicates that it properly detects the majority of positive cases. ViT makes fewer false positive predictions due to its extremely high precision. Compared to CNN, recall is somewhat poorer, meaning it overlooks more good cases. In every metric, the hybrid model performs the best.

Because of this, the entire row appears bright yellow, signifying an almost ideal value. When combined with the prior ROC curve (AUC = 1.00) and training performance (100% accuracy), the results jointly demonstrate the hybrid model's robustness and reliability in lung nodule classification. These high values reflect the model's stability and great generalization abilities in identifying both positive and negative lung nodules. The hybrid technique effectively reduces false positives and false negatives by balancing precision and memory, as evidenced by the high F1 score.

Table.1. Comparison study of the Hybrid method with its results

Author	Database	Hybrid method	Precision	Recall	F1	Accuracy
[7]	LIDC dataset	Hybrid WOA_APSO algorithm	-	-	-	97.5
[5]	IQ-OTH/NCCD - Lung Cancer Dataset Chest CT-Scan images Dataset	LwViT-DL	99.605	99.24	99.405	99.69
[10]	IQ-OTH/NCCD	InceptionNeXt-Based Hybrid	99.67	99.60	99.12	99.54
[13]	Lung and Colon Cancer Histopathological Images Dataset	ViT-DCNN	94.37	94.24	94.23	94.24
[12]	LC25000	FusionNet - ViT	99.29	98.69	99.04	99.36
[11]	LIDC-IDRI dataset	Hybrid CNN feature model	-	-	-	91.75
[14]	LIDC-IDRI dataset	Res-trans networks	91.62			92.92
Proposed model	LIDC-IDRI dataset	Hybrid CNN-ViT	99.45	99.45	99.45	99.43

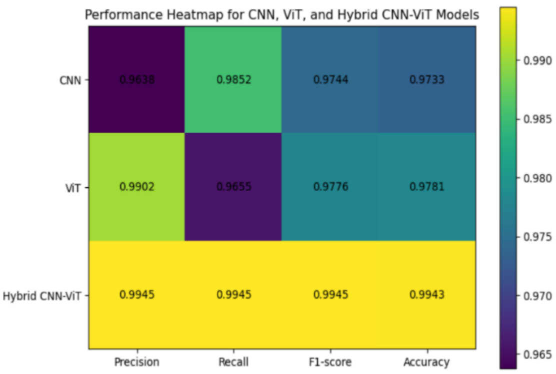
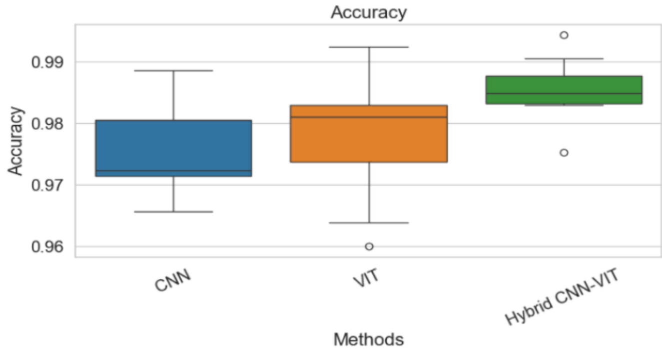
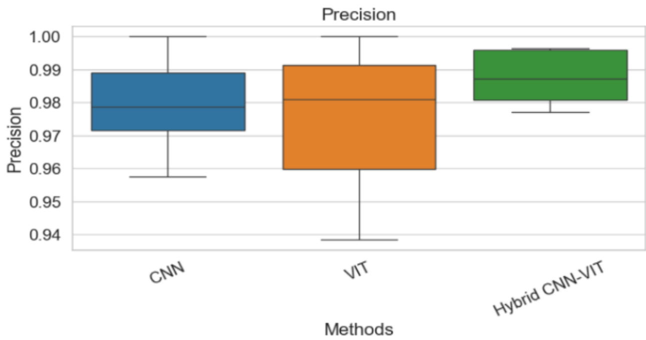


Fig.6. Heatmap of CNN, VIT and Hybrid CNN-VIT model with Precision, Recall, F1-score and Accuracy

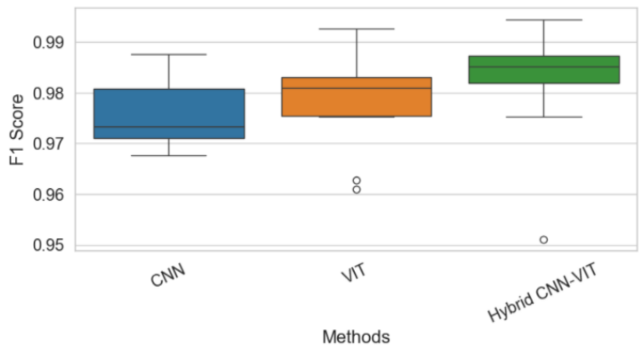
The Table.1 presents a comparison of the proposed Hybrid CNN–ViT model with current approaches. The proposed method achieves better or comparable performance based on precision, recall, F1-score, and accuracy, proving the efficacy of combining convolutional and transformer-based feature representations for lung nodule classification.



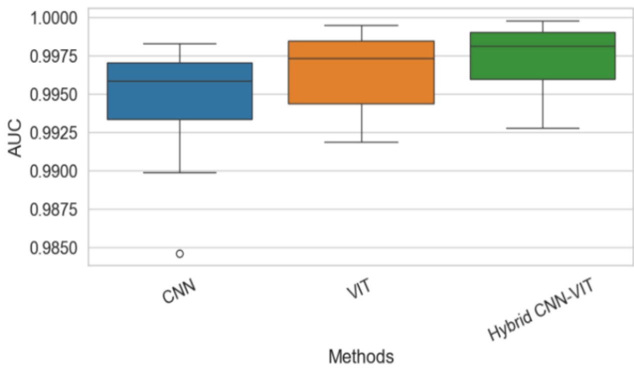
(a)



(b)



(c)



(d)

Fig.7. Comparison of statistical results (Accuracy, Precision, F1 Score, AUC) of CNN, VIT and Hybrid CNN-VIT.

The Fig.7(a) - Fig.7(d) provides strong evidence that the hybrid CNN-ViT model outperforms each baseline (CNN-only or ViT-only) on nearly all evaluated metrics: accuracy, precision, recall, F1, and AUC respectively, and does so in a stable, consistent way across experiments. For a lung-nodule classification task, this suggests the hybrid model is likely more reliable and clinically useful than either base model alone.

5. CONCLUSION

The best aspects of both architectures are combined to create a CNN-ViT hybrid model for lung nodule classification. The model's performance is promising and establishes a foundation for future studies on hybrid medical image analysis models. Future additions include adding multi-scale features and attention methods in fusion, as well as testing on larger, multi-class datasets. The proposed hybrid ResNet–ViT architecture exhibits remarkable promise for automated lung nodule classification with almost perfect metrics on all evaluation parameters. The outcomes show the model's effectiveness, stability, and adaptability in medical image processing tasks, paving the way for more study in useful diagnostic applications.

6. ACKNOWLEDGEMENT

The authors gratefully acknowledge the project funding received from Chief Minister's Research Grant (CMRG) Scheme, Government of Tamil Nadu. The authors are also thankful to Bharathiar University for valuable support.

7. FUNDING

This research work is funded under the Chief Minister's Research Grant Scheme, Directorate of Technical Education, Government of Tamil Nadu, India.

8. DATA AVAILABILITY STATEMENT

This article used the LIDC-IDRI public dataset (https://www.cancerimagingarchive.net/).

REFERENCES

[1] H.J.W.L. Aerts, L. Wee, E. Rios Velazquez, R.J. Gillies and P. Lambin, "Data From NSCLC-Radiomics (Version 4) [Data set]", Available at: https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI, Accessed in 2019.

[2] L.M. Li, H. Zhang and H.T. Yu, "Few-Shot Learning for CT Lung Nodule Detection Based on Open-Set Object Detection", *Current Medical Science*, Vol. 45, pp. 1358-1366, 2025.

[3] Data From LIDC-IDRI The Cancer Imaging Archive, Available at: https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX, Accessed in 2015.

[4] X. Rafael-Palou, A. Jimenez-Pastor and L. Marti-Bonmati, "Advancing Deep Learning-Based Segmentation for Multiple Lung Cancer Lesions in Real-World Multicenter CT Scans", *European Radiology*, Vol. 9, pp. 78-94, 2025.

[5] D. Mannepalli, Tan Kuan Tak, Sivaneasan Bala Krishnan and Velagapudi Sreenivas, "GSC-DVIT: A Vision Transformer based Deep Learning Model for Lung Cancer Classification in CT Images", *Biomedical Signal Processing and Control*, Vol. 103, pp. 107371-107385, 2025.

[6] Changyong Li., Y. Fan and X. Cai, "PyConvU-Net: A Lightweight and Multiscale Network for Biomedical Image Segmentation", *BMC Bioinformatics*, Vol. 22, No. 14, pp. 1-17, 2021.

[7] Vijh Surbhi, P. Gaurav and H.M. Pandey, "Hybrid Bio-Inspired Algorithm and Convolutional Neural Network for Automatic Lung Tumor Detection", *Neural Computer and Applications*, Vol. 35, pp. 23711-23724, 2023.

[8] Hui Yu, J. Li and L. Zhang, "Design of Lung Nodules Segmentation and Recognition Algorithm based on Deep Learning", *BMC Bioinformatics*, Vol. 22, No. 5, pp. 314-325, 2021.

[9] A. Gondkar, Raju Ramakrishna, Sureka R. Gondkar, S.S. Kavitha and R.V. Siva Balan, "Hybrid Deep Learning Based GRU Model for Classifying the Lung Cancer from CT Scan Images", *Proceedings of 3rd International Conference on Distributed Computing and Electrical Circuits and Electronics*, pp. 1-8, 2024.

[10] A. Ozdemir, Burhanettin, Emrah Aslan and Ishak Pacal, "Attention Enhanced Inception next based Hybrid Deep Learning Model for Lung Cancer Detection", *IEEE Access*, Vol. 13, pp. 27050-27069, 2025.

[11] T. Zhao, Huafeng Wang, Lihong Li, Yifang Qi, Haoqi Gao, FangFang Han, Zhengrong Liang, Yanmin Qi and Yuan Cao, "A Hybrid CNN Feature Model for Pulmonary Nodule Differentiation Task", *Proceedings of International Workshop on Bio-Imaging and Visualization for Patient-Customized Simulations*, pp. 19-26, 2017.

[12] O. Singh, "FusionNet-ViT: Hybrid Deep Learning Model for Lung Cancer Classification", *SN Computer Science*, Vol. 6, pp. 940-962, 2025.

[13] A. Pal, Hari Mohan Rai, Joon Yoo, Sang-Ryong Lee and Yooheon Park, "ViT-DCNN: Vision Transformer with Deformable CNN Model for Lung and Colon Cancer Detection", *Cancers*, Vol. 17, No. 18, pp. 1-26, 2025.