TransResNet: A Dual-Stream Deep Model for Precision Pulmonary Lesion Classification

Sanjuktarani Jena[1], Utsav Avaiya[2],Upendra Kumar[3],Sai Kiran Oruganti[4]
Sardar Patel Institute of Technology, Mumbai, India[1,2]
Dr A P J Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India[3]
Lincoln University College, Malaysia[4]
sanjuktarani.jena@spit.ac.in, utsav.avaiya@spit.ac.in, upendra.ietlko@gmail.com,saisharma@lincoln.edu.my

Abstract—Lung cancer is one of the major cause of death due to cancer across the globe, thus an early detection is imperative for improving the outcome for patients diagnosed with this disease. This paper introduces a novel hybrid deep learning model that combines Convolutional Neural Network (CNN) which is pretrained, specifically ResNet-50, with a Transformer encoder to classify lung cancer directly from chest CT scans of lung images. The ResNet-50 backbone is chosen for its efficient local feature extraction, while the Transformer encoder captures global contextual relationships among features using self-attention mechanisms. The model was trained on IQ-OTH/NCCD lung cancer dataset, which comprises of 1190 CT scan slices from 110 patients. Preliminary experimental results indicate that the hybrid model attains a classification accuracy of 98.3%, outperforming traditional CNN-based models. The integration of CNNs and Transformer architecture significantly enhances the model's ability to capture both spatial and semantic patterns in medical images. This approach holds a high potential for early and accurate lung cancer detection in clinical settings which could lead to a higher possibility of cure for patients.

Index Terms—Deep learning, ResNet-50, Transformer Encoder, Attention Mechanism, Convolutional Neural Networks (CNN), Transfer Learning, Computed Tomography (CT)

I. INTRODUCTION

Lung cancer has emerged as a major threat to the health care systems throughout the world. According to the World Health Organisation and the Global Cancer Observatory (GLOBOCAN 2024) there have been 2.3 million new cases of lung cancer annually out of which 1.85 million leads to death, which accounts to almost 18% of all cancer deaths worldwide. Even though there has been progress by leaps and bounds in the methods of diagnosis and therapy, the survival rate at the five-year mark has been a paltry 19% (2024) mostly due to late diagnosis and the lack of therapy at that point. Earlier detection and correct classification of the subtypes of lung cancer is imperative for better prognosis and treatment. Currently imaging based dataset like chest X-Rays and lose dose Computed Tomography (CT) scans are the main tools. These methods, though, cheap and accessible, rely on manual interpretation by radiologists which is subjective as well as time consuming, and has a high margin of error. This creates a need for an automated system which helps clinicians detect malignant nodules with more accuracy.

Deep learning and Machine learning have enhanced medical image analysis in the recent past by leaning on automatic features of learning. Conventional ML methods depend on features designed manually which are extracted from imaging shape, texture and intensity but are held back by the quality of these input. Algorithms like convolutional neural networks (CNNs) based on Deep learning have outperformed other methods by functioning based on hierarchical representations from raw imaging data, thus enhancing sensitivity, specificity and detection Recently more complex architectural models like Residual Networks (ResNet50), Darknet MobileNet, and Vision Transformers (ViT) have further enhanced lung cancer classification by using complex spatial and contextual information in CT scans. Vision Transformers, use self-attention mechanisms, can model long-range dependencies well and have demonstrated promising results for medical imaging tasks. Further, hybrid frameworks that combine CNNs with neural networks that are recurrent, like Long Short-Term Memory (LSTM) facilitate temporal analysis of sequential imaging data to improve detection of tumor progression over time.

This work investigates cutting-edge deep learning and machine learning methods for lung cancer detection and diagnosis with benchmark datasets like LIDC-IDRI and LUNA16. We compare and assess several models such as CNN variants, ResNet50, Darknet, MobileNet, Vision Transformers, and CNN-LSTM hybrids with an aim to enhance diagnostic precision, minimize false positives, and facilitate clinical decision-making.

Through the combination of advanced deep learning frameworks and vast medical imaging data, this work aims to facilitate an earlier and more accurate diagnosis of lung cancer, which finally adds on to a better outcome for the patient and a decreasing global disease burden in the coming years.

II. LITERATURE SURVEY

Recent advances in deep learning have significantly enhanced computer-aided diagnosis (CAD) systems for lung cancer diagnosis. Among these, Transformer-based and hybrid CNN-Transformer models have demonstrated notable improvements in accuracy, interpretability, and generalizability.

Sun *et al.* [1] proposed an improved Swin Transformer architecture for lung image classification and segmentation, achieving 98.14% accuracy on CT datasets. Their model outperformed traditional CNNs by effectively modelling long-range dependencies and spatial context. Similarly, Ko *et al.* [2] applied Vision Transformers (ViTs) to chest X-rays, achieving an AUC of 0.956, surpassing baseline CNN models like ResNet-50 and DenseNet-121.

Ali *et al.* [3] presented a comprehensive review of Transformer applications across imaging modalities, emphasizing ViT's generalizability. Gai *et al.* [4] found that Transformer-based models outperformed CNNs in sensitivity and specificity, particularly on imbalanced datasets. Mart´ın and Sanchez [5] extended ViT to multimodal tumor classification, ´ reporting 95.6% accuracy on lung, brain, and kidney datasets. Mazumder and Liu [6] designed a dual-stage ViT model that achieved 97.2% accuracy on the LUNA16 dataset, validating its utility in both classification and localization tasks.

Hybrid models also show promise. Kanipriya *et al.* [7] developed a CNN-LSTM model optimized using the Capuchin Search Algorithm, achieving 96.1% accuracy. Lu *et al.* [8] combined CNN and RNN to predict survival outcomes from low-dose CTs with a concordance index of 0.82. Shafi and Chinnappan [9] proposed a CNN-Transformer-LSTM model yielding 98.3% accuracy and a Dice score of 97.5%. Kesiku and Garcia-Zapirain [10] introduced a hybrid AI model that achieved 97.5% accuracy, excelling at early-stage tumor detection.

BiLSTM networks have been utilized for their temporal modelling. Diao *et al.* [11] fine-tuned a BiLSTM with attention to classify nodules at 96.7% accuracy. Indumathi and Siva [12] proposed a CNN-BiLSTM with multi-head attention, achieving 97.1% accuracy and 96.3% sensitivity. Li *et al.* [13] developed a CNN-BiLSTM hybrid model for respiratory disease diagnosis, reaching 94.6% accuracy with enhanced noise robustness.

Traditional CNNs continue to perform well. Mamun *et al.* [14] developed LCDCTCNN on low-dose CT images, achieving 95.5% accuracy. Saha *et al.* [15] proposed VER-Net, a transfer learning-based hybrid model with 97.6% accuracy and an AUC of 0.984. Kalkan *et al.* [16] conducted a comparative analysis of CNN methods for lung cancer classification and identified CT-specific advantages in performance.

Ensemble and optimization methods have further pushed performance boundaries. Kumaran *et al.* [18] used an ensemble of VGG16, ResNet-50, and InceptionV3 with Grad-CAM to achieve 98.1% accuracy. Pandit *et al.* [19] enhanced CNNs through an optimized deep learning pipeline, reaching 96.8% accuracy. Li *et al.* [20] integrated machine learning classifiers with tuned hybrid features to achieve 97.3% accuracy. Mohamed *et al.* [21] employed the Ebola Optimization Search Algorithm, achieving 97.9% classification accuracy.

Transformer-hybrid models consistently achieve high performance. Durgam *et al.* [22] proposed a CNN-Transformer model that attained 97.84% accuracy and a 97.61% F1-score on CT scans. Gulsoy and Kablan [23] validated ViTs on CT-based diagnosis, reporting an F1-score of 97.67% and improved interpretability via attention maps. Wani *et al.* [26] introduced DeepXplainer, an interpretable Transformer-based model achieving 98.4% classification accuracy.

III. RESEARCH AND METHODOLOGIES

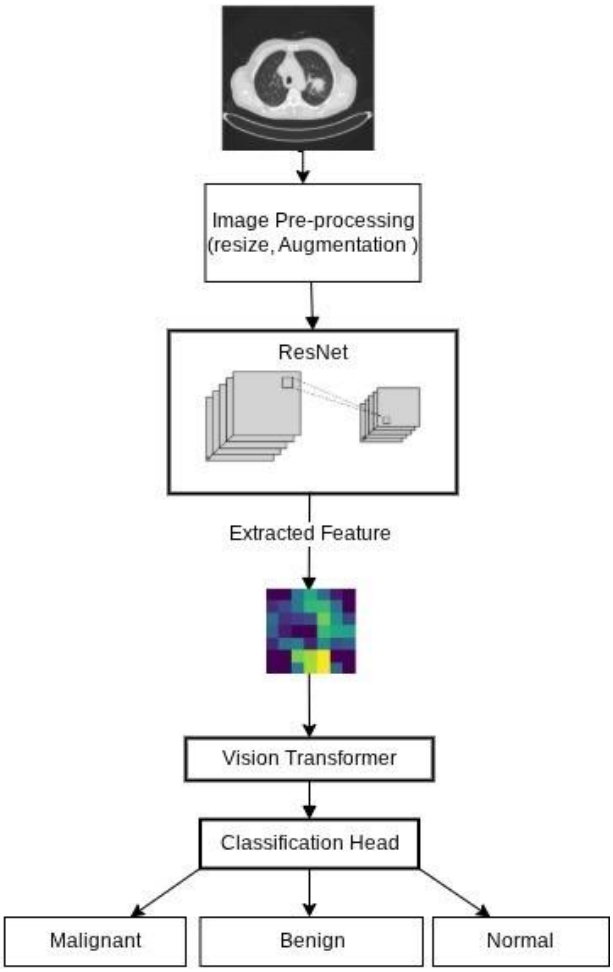

Fig. 1: Proposed Block Diagram

The proposed approach leverages a hybrid deep learning architecture that integrates a ResNet-50 convolutional backbone with a Transformer encoder to perform lung cancer diagnosis from CT scan images. The rationale behind this hybrid design is rooted in the complementary strengths of convolutional neural networks (CNNs) and Transformers. CNNs are proficient at capturing local patterns such as textures, edges, and small-scale structural variations within lung tissue. However, they are inherently limited by their local receptive fields and struggle to model long-range dependencies, which are crucial in medical imaging tasks where discriminative features may be spatially dispersed across the image.

Transformers, by contrast, employ self-attention mechanisms that allow the model to consider relationships between all spatial regions simultaneously, enabling better understanding of global context. This makes them particularly useful for interpreting complex patterns in medical images where spatial coherence and relational cues play a significant role. In the proposed pipeline, a pre-trained ResNet-50 (adapted to accept single-channel grayscale inputs) is used to extract spatial feature maps from CT slices resized to 224 × 224 pixels. The final feature maps from the CNN are reshaped into a sequence of 49 feature tokens, each corresponding to a 7 × 7 spatial patches. These tokens are linearly projected and enriched with positional encodings before being passed to a Transformer encoder comprising two layers with multi-head self-attention.

The output sequence is aggregated using mean pooling and passed through a layer normalization and a fully connected layer to produce the final class logits. To bridge the gap created by the class imbalance inherent in the dataset, the Focal Loss is employed, which down-weights well-classified examples and focuses learning on harder, misclassified ones. The Focal Loss is defined as:

$$L_{focal} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \qquad (1)$$

where p_t is the model's estimated probability for the true class t, α_t is the weighting factor for class t, and γ is a tunable focusing parameter that emphasizes harder examples. In our setup, γ was set to 2, and α_t was computed inversely proportional to class frequencies.

To evaluate the model, a stratified 5-fold cross-validation strategy was implemented, ensuring that class proportions were maintained across splits. For each fold, the model was trained for up to 50 epochs incorporating Adam optimizer. The total number of trainable parameters are 25,619,651.

IV. DATASET AND EXPERIMENTAL SETUP

The experiment was conducted on the IQ-OTH/NCCD lung cancer dataset, which comprises of 1190 Computed Tomography scan slices from 110 patients, categorized into classes as malignant, benign, and normal. The data was collected from two major cancer hospitals in Iraq during a three-month period in 2019. Each patient case contains multiple CT slices (ranging from 80 to 200), providing varied views of the chest region. The scans were originally in DICOM format and acquired using a Siemens SOMATOM CT scanner. The dataset reflects real-world class imbalance, with 40 cases being malignant, 15 being benign, and the rest 55 being normal cases. This makes it well-suited for evaluating classification models under practical, imbalanced conditions.

All CT slices were converted to grayscale and rescaled to 224 × 224 pixels. Techniques like random rotation and horizontal flipping were applied for data augmentation to improve generalization, along with resizing all images to a fixed resolution. Proposed method used a stratified 5-fold cross-validation approach to assess model performance across different splits while handling class imbalance using Focal Loss and weighted sampling. Among the five trained models, the one with the maximum validation accuracy was selected as the best model, and all reported results correspond to this fold. The proposed model combines a ResNet-50 backbone for spatial feature extraction with a Transformer encoder for modelling long-range dependencies. The model was trained with a batch size of 64, for 50 epochs optimized by the Adam optimizer with an initial learning rate of 10^{-4}, and a ReduceLROnPlateau scheduler. All experiments were conducted on a Kaggle Notebook utilizing the free NVIDIA Tesla P100 GPU runtime.

V. RESULT AND DISCUSSION

The proposed hybrid model, which integrates ResNet-50 with a Transformer encoder, achieved a validation accuracy of 98.63%, demonstrating strong performance across all classes of lung cancer. Notably, the model exhibited excellent precision and recall for the malignant class, achieving an F1score of 0.99, which highlights its robustness in detecting the most critical cases. The benign class, despite having a smaller number of samples, achieved an F1-score of 0.97, showing the model's capability to generalize well even with limited data. The normal class also recorded a high F1score of 0.99. The overall macro-average F1-score of 0.98 and weighted-average F1-score of 0.99 reflect the model's balanced performance across all categories. These findings confirm the effectiveness of combining CNN and Transformer architectures in capturing both spatial and contextual patterns, resulting in precise classification even under class imbalance conditions.

TABLE I: Final Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.98	1.00	0.99	84
Benign	1.00	0.94	0.97	16
Malignant	0.99	0.98	0.99	119
Accuracy			0.99	219
Macro Avg	0.99	0.97	0.98	219
Weighted Avg	0.99	0.99	0.99	219

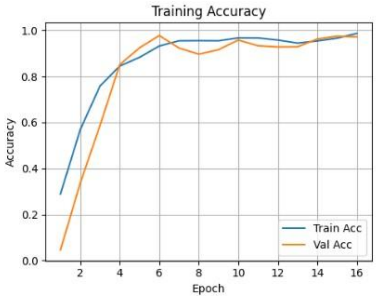

Fig. 2: Training and Validation Accuracy

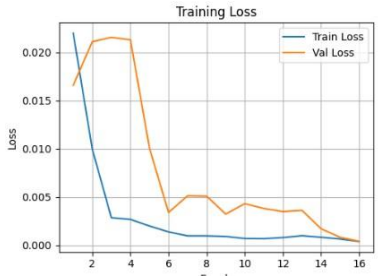

Fig. 3: Training and Validation Loss

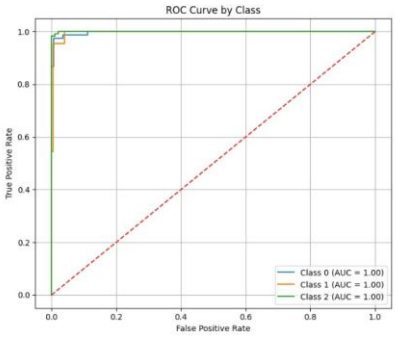

Fig. 4: ROC Curve

VI. CONCLUSION

This study presents a hybrid model built upon deep learning, that combines ResNet-50 and a Transformer encoder for classification of lung cancer from chest CT scans images. The model effectively captures both local and global features, outperforming traditional CNN approaches and achieving an accuracy of 98.63%. By reducing the dependency on manual interpretation, it offers a reliable tool for early and accurate detection. With further validation, this model can strengthen and amplify the pace as well as accuracy in the decision-making process of the clinician and improve overall outcomes.

REFERENCES

[1] R. Sun, Y. Pang, and W. Li, “Efficient lung cancer image classification and segmentation algorithm based on an improved Swin Transformer,” *Electronics*, vol. 12, no. 4, p. 1024, 2023.

[2] J. Ko, S. Park, and H. G. Woo, “Optimization of vision transformerbased detection of lung diseases from chest X-ray images,” *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 191, 2024.

[3] H. Ali, F. Mohsen, and Z. Shah, “Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review,” *BMC Medical Imaging*, vol. 23, no. 1, p. 129, 2023.

[4] L. Gai, M. Xing, W. Chen, Y. Zhang, and X. Qiao, “Comparing CNNbased and transformer-based models for identifying lung cancer: which is more effective?,” *Multimedia Tools and Applications*, vol. 83, no. 20, pp. 59253–59269, 2024.

[5] O. A. Mart´ ´ın and J. Sanchez, "Evaluation of Vision Transformers for´ Multimodal Image Classification: A Case Study on Brain, Lung, and Kidney Tumors," *arXiv preprint arXiv:2502.05517*, 2025.

[6] A. Mazumder and J. Liu, "Developing a dual-stage vision transformer model for lung disease classification," *arXiv preprint arXiv:2409.18257*, 2024.

[7] M. Kanipriya, C. Hemalatha, N. Sridevi, S. R. SriVidhya, and S. J. Shabu, "An improved capuchin search algorithm optimized hybrid CNN-LSTM architecture for malignant lung nodule detection," *Biomedical Signal Processing and Control*, vol. 78, p. 103973, 2022.

[8] Y. Lu *et al.*, "A hybrid CNN-RNN approach for survival analysis in a Lung Cancer Screening study," *Heliyon*, vol. 9, no. 8, 2023.

[9] S. M. Shafi and S. K. Chinnappan, "Hybrid transformer-CNN and LSTM model for lung disease segmentation and classification," *PeerJ Computer Science*, vol. 10, p. e2444, 2024.

[10] C. Y. Kesiku and B. Garcia-Zapirain, "AI-Enhanced Lung Cancer Prediction: A Hybrid Model's Precision Triumph," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[11] S. Diao *et al.*, "Optimizing Bi-LSTM networks for improved lung cancer detection accuracy," *PLOS ONE*, vol. 20, no. 2, p. e0316136, 2025.

[12] V. Indumathi and R. Siva, "Improving Early Detection of Lung Disorders: A Multi-head Self-Attention CNN-BiLSTM Model," *Journal of The Institution of Engineers (India): Series B*, vol. 105, no. 3, pp. 595– 607, 2024.

[13] L. Li *et al.*, "Prediction and Diagnosis of respiratory disease by combining convolutional neural network and bi-directional long shortterm memory methods," *Frontiers in Public Health*, vol. 10, p. 881234, 2022.

[14] M. Mamun, M. I. Mahmud, M. Meherin, and A. Abdelgawad, "LCDCTCNN: Lung cancer diagnosis of CT scan images using CNN-based model," in *Proc. 10th Int. Conf. Signal Processing and Integrated Networks (SPIN)*, pp. 205–212, IEEE, 2023.

[15] A. Saha *et al.*, "VER-Net: a hybrid transfer learning model for lung cancer detection using CT scan images," *BMC Medical Imaging*, vol. 24, no. 1, p. 120, 2024.

[16] M. Kalkan *et al.*, "Comparative Analysis of Deep Learning Methods on CT Images for Lung Cancer Specification," *Cancers*, vol. 16, no. 19, p. 3321, 2024.

[17] A. Shahzad *et al.*, "Pneumonia Classification from Chest X-ray Images Using Pre-Trained Network Architectures," *VAWKUM Transactions on Computer Sciences*, vol. 10, no. 2, pp. 34–44, 2022.

[18] S. Y. Kumaran *et al.*, "Explainable lung cancer classification with ensemble transfer learning of VGG16, ResNet50 and InceptionV3 using Grad-CAM," *BMC Medical Imaging*, vol. 24, no. 1, p. 176, 2024.

[19] B. R. Pandit *et al.*, "Deep learning neural network for lung cancer classification: enhanced optimization function," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 6605–6624, 2023.

[20] L. Li *et al.*, "Enhancing lung cancer detection through hybrid features and machine learning hyperparameters optimization techniques," *Heliyon*, vol. 10, no. 4, 2024.

[21] T. I. Mohamed, O. N. Oyelade, and A. E. Ezugwu, "Automatic detection and classification of lung cancer CT scans based on deep learning and Ebola optimization search algorithm," *PLOS ONE*, vol. 18, no. 8, p. e0285796, 2023.

[22] R. Durgam *et al.*, "Enhancing lung cancer detection through integrated deep learning and transformer models," *Scientific Reports*, vol. 15, no. 1, p. 15614, 2025.

[23] T. Gulsoy and E. B. Kablan, "Diagnosis of lung cancer based on CT scans using Vision Transformers," in *Proc. 14th Int. Conf. Electrical and Electronics Engineering (ELECO)*, pp. 1–5, IEEE, 2023.

[24] K. T. Chui *et al.*, "Multiround transfer learning and modified generative adversarial network for lung cancer detection," *International Journal of Intelligent Systems*, vol. 2023, Article ID 6376275.

[25] M. Nishio *et al.*, "Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning," *PLOS ONE*, vol. 13, no. 7, p. e0200721, 2018.

[26] N. A. Wani, R. Kumar, and J. Bedi, "DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence," *Computer Methods and Programs in Biomedicine*, vol. 243, p. 107879, 2024.