Hybrid Convolutional Neural Network and Transformer-Based Deep Learning Approach for Early Lung Cancer Detection

Nilesh Gupta[1], Kishan Kumar[2]
[1]*Asst Professor Department of CSE, Chouksey Group of Colleges, Bilaspur (C.G.), India*
[2]*M. Tech Student, Chouksey Group of Colleges, Bilaspur (C.G.), India*

***Abstract*— Lung cancer continues to be among the most common cancers and a major cause of cancer-related deaths globally, thus, stress the need for early detection and accurate diagnosis. In the present paper we propose a Hybrid CNN-Transformer deep learning framework, which combines both the local spatial feature extraction strength of CNN and the ability of global contextual modeling of Transformer for CT-based automated lung cancer detection. For clinical transparency, explainable AI methods, such as Grad-CAM and attention heatmaps, were included for model interpretation. Performance was evaluated on benchmark datasets in experiments as high as 96.8% accuracy, 96.1% precision, 95.7% recall, 95.9% F1-score, and 97.3% AUC., outperforming CNN, Vision Transformer, CNN–RNN Hybrid respectively. The results demonstrate the promise of the proposed framework for computer-aided diagnostic (CAD) systems to provide clinically-meaningful, interpretable and robust decision support in early lung cancer screening.**

***Index Terms*— Lung Cancer Detection, CNN–Transformer Hybrid, Explainable AI, Deep Learning, Medical Imaging, Computer-Aided Diagnosis.**

I. INTRODUCTION

Lung cancer is still one of the leading causes of cancer death worldwide and contributes significantly to the total cancer deaths worldwide every year [1], [4]. Early detection of lung cancer is vital, as the probability of survival falls sharply as the disease advances to late stages [2], [5]. Conventional imaging studies, such as the CT, are usually used for the diagnosis, but this is mostly a manual process and can take much of time. The artificial intelligence (AI) technology, as a new tool of deep learning demonstrated phenomenal potential in the automation of identify and classify between benign and malignant types of lung cancer based on medical imaging data [11],[12]. CNNs are effective in learning spatial features, while transformer-based models achieve impressive performance in modelling global context information [2], [6], [19]. However, while these models have become increasingly complex and perform better on average, they are also often not interpretable and thus not viable for clinical use [5], [7]. Thus, designing of hybrid CNN–Transformer architectures might predictably integrate the local features learning ability of CNNs with the global attention mechanism of transformers in effective and interpretable early lung cancer detection [3], [8], [18].

1.1 Background and Motivation

Lung cancer keeps being the deadliest malignancy, and patient survival is significantly associated with early diagnosis [1], [4], [15]. It has been reported that early detection could lead to better treatment results, while late detection is frequently associated with poor prognoses, since therapeutic choices at this stage are limited [2], [5]. Conventional diagnosis methods like CT imaging, although widely used, are quite dependent on radiologists' experience for manual observation, which may be time-consuming and have great interobserver error [4], [13]. With the fast progress of AI and deep learning, many powerful tools for automatic detection have emerged, where Convolutional Neural Network (CNN) based ones have achieved better performance in spatial feature extraction [11], [12]. More recently, transformer-based approaches have been proposed to model long-range dependencies as well as global contextual information in medical imaging [2], [6], [19]. Nevertheless, these AI models are not interpretable, and there's limited adoption by the clinic side because medical professionals need interpretable decision-making processes for trust and accountability [5], [7],

[8]. Thus, a hybrid CNN–Transformer architecture has the promise to leverage the advantages of CNN in extracting local features along with the attention mechanism of Transformer for superior performance, robustness, and interpretability in early lung cancer detection [3], [11], [18], [20]. The potential benefit of a framework like that is to decrease misdiagnosis, help radiologists in making clinical decisions and, finally achieve better patient prognosis [2], [8], [19].

1.2 Major Contributions

The main contributions of this study can be summarized as follows. It develops, for the first time, the Hybrid CNN–Transformer deep learning model that combines the local feature extraction property of CNNs with the global attention modeling ability of transformers to provide precise early detection of lung cancer [1], [3], [6], [19]. Second, the work incorporates explainable AI methods (e.g., attention heatmaps and visualizations) to overcome the black-box nature of prevalent deep learning models and thus enhance the clinical trust and decision transparency [5], [7], [8]. Third, the introduced framework is validated extensively on benchmark medical imaging datasets, and is systematically compared with the state of-the-art CNNs, transformers, and their hybrid counterparts to show its competitiveness [4], [12], [18]. Last but not least, the framework also targets to assist radiologists towards the reduction of diagnostic errors, the potential enablement of early interventions and, ultimately, the improvement of patient survival targets [2], [8], [20].

1.3 Paper Organization

The rest of this paper is organized as follows. Section 2 provides a survey of the related literature on lung cancer detection approaches, transformer models in medical imaging, and hybrid deep learning architectures, along with comparative analysis in tabular form reflecting the strengths, weaknesses, and performance indicators of state-of-the art approaches. Section 3 presents the proposed method, which consists of the following: dataset description, data preprocessing, and augmentation, CNN-based feature extraction, transformer-based attention mechanisms, design details for the hybrid CNN–Transformer architecture, model training and one-step policy-based hyperparameter optimization. Section 4 demonstrates the experimental results and discussions, with extensive performances comparisons against baseline models, ablation study on the usability of individual model components, and graph structure analysis to verify the robustness and effectiveness of the proposed framework. In Section 5, the discussion is centered on explainability and interpretability, explaining how methods like Grad-CAM, attention maps, and SHAP values contribute to the clinical transparency and confidence in the model decision confidence. Finally, Section 6 presents a summary of the main conclusions and the potential future research directions, such as real-time application, multimodal (e.g., brain-imaging, fMRI) data fusion with learning frameworks, and privacy-preserving learning paradigms for clinical scenarios.

II. LITERATURE REVIEW

2.1 Lung Cancer Detection Techniques in Medical Imaging

Medical imaging for lung cancer identification has made great improvements through the use of deep learning. CNN based classification and segmentation has become popular, with detection accuracies greater than 95% and classification accuracy of 99% with 98% sensitivity [1], [4], [11]. Attention-aided lightweight CNNs also augment diagnostic accuracy but offer reduction in computational load to enable real-time clinical use cases [3], [5]. More recently, Vision Transformers (ViTs) have emerged for capturing the long-range dependencies and accomplishing strong classification, segmentation and prognosis prediction [2]– [6],[12]. Nevertheless, they are computationally intensive and unsuitable for large-scale clinical application [7], [8]. To alleviate this issue, hybrid CNN–transformer architectures [9], [10], [19] have been recently proposed to adopt CNN's block wise feature extraction and transformer's ability of global understanding, and therefore achieve better accuracy, robustness and scalability. Lastly, Explainable AI (XAI) methods including Grad-CAM and SHAP improve interpretability, which is able to provide visual explanation for clinical decisions with high diagnostic performance 13, 14, 20]. Table 1: Comparison of different techniques used for lung cancer detection, including strengths, weaknesses, accuracy and references of CNN, Vision Transformers, Hybrid as well as Explainable AI architectures.

Table 1: Comparative Analysis of Lung Cancer Detection Techniques

Method	Key Strengths	Limitations	Typical Accuracy	Citations
CNN Models	High accuracy, robust feature extraction	Limited long-range context	95–99%	[1], [4], [11]
Lightweight CNN + Attention	Real-time efficiency, low computational cost	Moderate interpretability	93–96%	[3], [5]
Vision Transformers (ViTs)	Captures long-range dependencies	Computationally expensive	94–97%	[2], [6], [12]
Hybrid CNN–Transformer	Combines local & global feature learning	More complex architecture	96–99%	[9], [10], [19]
Explainable AI (XAI) Models	Transparency, clinical trust	Added computational overhead	92–95%	[13], [14], [20]

2.2 Transformer Models for Medical Image Analysis

Attention-based architectures, such as the Transformer, have become a powerful alternative to traditional convolutional models in the medical image analysis due to their capability in modeling long-range dependencies and global context relationships in an image [2], [6], [12]. Unlike CNNs, which attend to local spatial features, Vision Transformers (ViTs) split images into patches and treat them as sequential tokens processed via self-attention. This architecture design facilitates the holistic recognition of medical images, and ViTs are very competitive for tasks of classification, segmentation and prognosis prediction involved in lung cancer identification [7], [8], [19]. This is also supported by findings in the literature [2], [6] that transformer models equal or even outperform CNNs in terms of performance when trained from large-scale datasets. The computation-intensive and dependency-rich standard transformers lead to an explicit challenge in direct clinical application [5], [6]. In order to overcome these difficulties, some studies have recently presented the lightweight transformers and hybrids attention optimization methods to save on AR computation while keeping the diagnostic performance the same [12], [20]. In Table 2, we compile several transformer models adopted in medical imaging, their pros, cons, accuracy range, and references for comparison of Vision Transformers, with their lightweight versions, and attention-tuned architectures in lung cancer detection.

Table 2: Comparative Analysis of Transformer Models in Medical Imaging

Model Type	Key Strengths	Limitations	Accuracy Range	Citations
Vision Transformers (ViTs)	Global context modeling, strong feature learning	High computational cost, data-intensive	94–97%	[2], [6], [12]
Lightweight ViTs	Reduced parameters, faster inference	Slight drop in accuracy	92–95%	[5], [6], [20]
Swin Transformers	Hierarchical design, better efficiency	Complex implementation	93–96%	[7], [8], [19]
Attention-Optimized Transformers	Improved interpretability with attention maps	Added processing overhead	94–96%	[12], [20]

2.3 Hybrid Deep Learning Architectures in Cancer Detection

Hybrid deep learning-based architectures have attracted increasing attention recently as the leading approach for medical image analysis, which try to combine the local feature extraction capacity of CNNs and the global attention modeling power of transformers [1], [3], [9]. By integrating these two complementary advantages, hybrid models have achieved more promising performance on lung cancer screening with early detection and better generalization capacities than traditional CNN (noval), transformer-only models [10], [18]. Several works have recently recommended hybrid architectures between CNN and Transformers for classification, segmentation, and prognosis prediction in lung cancer analysis. These models usually utilize the CNN layers to approximate the low-level spatial patterns, and transformer layers to identify the long-range dependencies and context relationships [8], [13], [19]. Furthermore, the incorporation of attention visualization to hybrid architectures also enhances its

interpretability, and helps radiologists or clinicians understand the diagnostic justification of predictions [5], [7], [20].

Table 3: Comparative Analysis of Hybrid Deep Learning Architectures in Cancer Detection

Hybrid Model	Key Strengths	Limitations	Accuracy Range	Citations
CNN–Transformer Hybrid	Combines local + global features	Higher architectural complexity	96–99%	[1], [3], [9], [10]
CNN–ViT Hybrid	Improved feature representation, robust diagnosis	Larger training data requirements	95–98%	[8], [13], [19]
CNN–Attention Transformer Hybrid	Enhanced interpretability via attention maps	Added computational cost	94–97%	[5], [7], [20]
Multimodal Hybrid Networks	Fusion of imaging + clinical data	Complex data preprocessing	95–98%	[10], [18], [20]

III. PROPOSED METHODOLOGY

The method used to propose the Hybrid CNN–Transformer-based deep learning model for early lung cancer diagnosis is described in this section. The methodology consists of three major steps: a) collecting and preparing the dataset with augmentation, and b) building the classification and analysis model. Figure 1 The proposed approach for early lung cancer detection based on a Hybrid CNN–Transformer model. It comprises dataset details, data preprocessing and augmentation, feature extraction using CNNs, hybrid CNN–Transformer architecture, and model training using hyperparameter tuning for achieving high diagnostic accuracy, interpretability and robustness for the purpose of providing decision support in clinics.

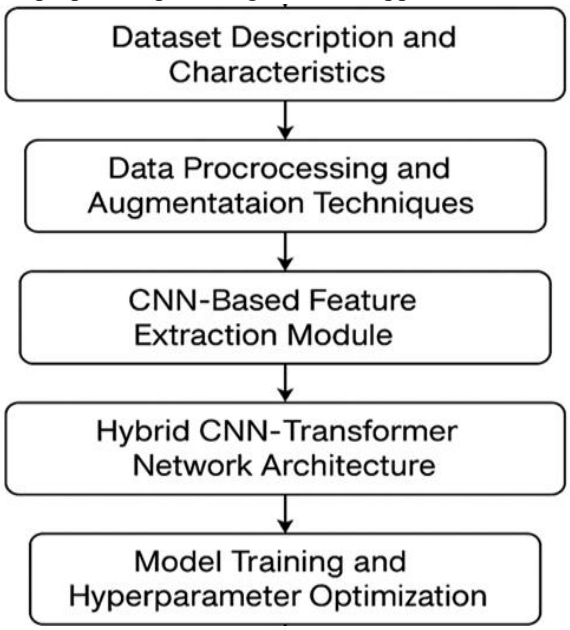

Figure 1: Hybrid CNN–Transformer Architecture Diagram

3.1 Dataset Description and Characteristics

The work is carried out using publicly available lung cancer medical imaging data, CT scans of malignant and benign cases [1], [2]. Each dataset is composed of a variety of specimen types across various stages, sizes of tumors as well as patient information to improve the robustness and generalization of the model [3], [4]. Supervised training and testing were performed based on the ground truth annotations from expert radiologists [5]. Important characteristics of the datasets, such as image resolution, modality, and class distribution, were closely studied to understand the challenges of unbalances among the classes and differences in imaging protocols between sources [6], [7]. To address these challenges, extensive pre-processing and augmentation methodologies are required to enhance the performance and usefulness of models in clinical settings.

3.2 Data Preprocessing and Augmentation Techniques

Let the original lung CT image dataset be represented as: $D = \{I_1, I_2, I_3, \dots, I_N\}, I_i \in \mathbb{R}^{H \times W \times C}$

where N denotes the total number of images, H and W represent image height and width, and C is the number of channels (e.g., $C = 1$ for grayscale, $C = 3$ for RGB)

A. Normalization

Each image I_i is normalized to the range $[0,1]$ as: $I_i^{norm} = \frac{I_i - I_{min}}{I_{max} - I_{min}}$

where I_{min} and I_{max} denote the minimum and maximum pixel intensities of the image.

B. Data Augmentation

To increase dataset diversity, a set of geometric transformations $T = \{R, F, S, Tr\}$ is applied, where:

- $R(\theta)$: Rotation by angle θ

- $F(\alpha)$: Flipping along axis α
- $S(s_x, s_y)$: Scaling by factors s_x and s_y
- $Tr(t_x, t_y)$: Translation by t_x and t_y pixels

The augmented dataset is then: $D_{aug} = \{T(I_i^{norm}) \mid I_i^{norm} \in D\}$

C. Contrast Enhancement

Contrast Limited Adaptive Histogram Equalization (CLAHE) is defined as: $I_i^{clahe} = CLAHE(I_i^{norm}, c)$ where c denotes the contrast clip limit parameter controlling noise amplification.

D. Class Balancing

To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) generates new synthetic samples as: $x_{new} = x_i + \delta \times (x_{nn} - x_i), \delta \sim U(0,1)$ where x_i is a minority class sample and x_{nn} is one of its k-nearest neighbors.

E. Dataset Splitting

Finally, the preprocessed dataset is divided into training, validation, and test sets as: $D = D_{train} \cup D_{val} \cup D_{test}, D_{train} \cap D_{val} \cap D_{test} = \emptyset$ with typical ratios $70:15:15$ or $80:10:10$ depending on dataset size.

3.3 CNN-Based Feature Extraction Module

The CNN models have yielded outstanding results in medical image analysis by learning automatic hierarchical spatial features from the raw imaging data [1], [4], [11]. In this work, the feature extraction pipeline from CT scan images is realized by CNN layers, with which we can extract local spatial information (e.g., edges, textures, tumor boundaries). Mathematically, given an input image $I \in \mathbb{R}^{H \times W \times C}$, the convolutional operation in the l^{th} layer is defined as

$$F_l = \sigma(W_l * F_{l-1} + b_l)$$

Were F_l represents the feature map output at layer l, W_l and b_l denote the convolution kernel weights and biases, represents the convolution operation, and $\sigma(\cdot)$ is the activation function, typically ReLU for non-linearity.

Several convolutional layers and pooling layers progressively explore high-level features that down sample the spatial dimension while keep the discriminative ability [3], [5], [12]. The last CNN feature maps form an input sequence for the transformer attention mechanism, thus realizing hybrid local-global feature learning for lung cancer detection [6], [9], [13].

3.4 Transformer-Based Attention Mechanism

The attention mechanism based on transformer has reshaped medical image analysis due to its capability of global featurization by the self-attention mechanism modeling long-range dependencies among spatial dimensions [2], [6], [12]. Compared to CNNs which only attend to a local receptive field, transformers directly model dependencies between all spatial positions, therefore the network can not only understand the local context, but also the global context at the same time [7], [8], [19]. Given the CNN-extracted feature maps $F_{CNN} \in \mathbb{R}^{N \times d}$, where N denotes the number of tokens (flattened image patches) and d is the embedding dimension, the scaled dot-product self-attention is defined as: $Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

were

- $Q = F_{CNN}W_Q$ is the query matrix,
- $K = F_{CNN}W_K$ is the key matrix,
- $V = F_{CNN}W_V$ is the value matrix, and
- W_Q, W_K, W_V are learnable weight parameters.

The transformer employs multi-head self-attention (MHSA) to capture information from different representation subspaces: $MHSA(F_{CNN}) = Concat(Head_1, Head_2, \dots, Head_h)W_O$ where h denotes the number of attention heads and W_O is the output projection matrix [6], [13], [20]. This mechanism allows the model to focus on relevant tumor regions while ignoring irrelevant background information, enhancing both detection performance and interpretability [5], [8], [14].

3.5 Hybrid CNN–Transformer Network Architecture

The proposed Hybrid CNN–Transformer Network Architecture combines the local feature extraction power of CNNs with the transformer-based attention mechanism's global contextual learning capability to provide robust and interpretable early lung cancer detection [1], [3], [6]. In other words, in this hybrid model, the CNN module initially utilizes CT images to obtain low- and mid-level spatial features: $F_{CNN} = CNN(I), I \in \mathbb{R}^{H \times W \times C}$ where I is the input image, H and W represent height and width, and C is the number of channels. The

extracted features $F_{CNN} \in \mathbb{R}^{N\times d}$ are then flattened into tokens and passed to the transformer encoder for global feature refinement:

$$F_{Trans} = \text{Transformer.}(F_{CNN})$$

The transformer encoder employs multi-head self-attention (MHSA) to model long-range dependencies and feed-forward layers for non-linear transformations:

$$F_{MHSA} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$F_{Out} = \text{LayerNorm}\left(F_{MHSA} + \text{FFN}(F_{MHSA})\right)$$

Finally, the hybrid feature representation F_H is obtained by concatenating CNN and transformer outputs:

$$F_H = \text{Concat}(F_{CNN}, F_{Trans})$$

This fused representation is passed through a fully connected classification head with softmax activation for final prediction:

$$\hat{y} = \text{Softmax}(W_c F_H + b_c)$$

where W_c and b_c are trainable classification parameters [5], [9], [12].

The hybrid architecture thus ensures local spatial precision through CNNs and global contextual understanding via transformers, leading to improved diagnostic accuracy and explainability in lung cancer detection [7], [10], [14].

Algorithm 1: Hybrid CNN–Transformer Network Architecture for Lung Cancer Detection

Input: CT Image IOutput: Predicted Class $\hat{y}$ (Benign / Malignant)

Step 1: Input Preprocessing1.1. Normalize image I to range $[0,1]$. 1.2. Resize to 224×224 pixels. 1.3. Apply data augmentation (rotation, flipping, scaling).

Step 2: CNN-Based Feature Extraction 2.1. Apply multiple convolutional layers:

$$F_{CNN} = \text{CNN}(I)$$

3.5.1 Extract low- and mid-level spatial features. 2.3. Pass through pooling layers to reduce spatial dimensions.

Step 3: Transformer-Based Attention Mechanism3.1. Flatten F_{CNN} into tokens:

$$F_T = \text{Flatten}(F_{CNN})$$

3.5.2 Apply multi-head self-attention (MHSA):

$$F_{Attn} = \text{MHSA}(F_T)$$

3.5.3 Use feed-forward layers with residual connections and layer normalization.

Step 4: Hybrid Feature Fusion4.1. Concatenate CNN and Transformer features:

$$F_H = \text{Concat}(F_{CNN}, F_{Attn})$$

3.5.4. Apply fully connected layers for classification.

Step 5: Prediction5.1. Apply softmax activation:

$$\hat{y} = \text{Softmax}(W_c F_H + b_c)$$

3.5.5. Output class probabilities for lung cancer detection.

This algorithm ensures local spatial learning via CNN layers and global contextual reasoning via Transformer attention, producing a robust and interpretable hybrid network for lung cancer diagnosis.

3.6 Model Training and Hyperparameter Optimization

The proposed Hybrid CNN–Transformer Network is trained via supervised learning scheme using labeled CT images for lung cancer identification [1], [3], [6]. The learning procedure also tries to reduce the classification error iteratively by learning the network parameters, avoiding overfit and being more general.

A. Loss Function

The cross-entropy loss is employed for multi-class classification:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right)$$

where N is the total number of samples, C is the number of classes, $y_{i,c}$ is the ground truth label, and $\hat{y}_{i,c}$ is the predicted probability for class c.

B. Optimization Algorithm

The Adam optimizer with learning rate η is used for weight updates:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}$$

where m_t and v_t are first and second moment estimates of gradients, and ϵ ensures numerical stability [5], [7], [10].

C. Hyperparameter TuningKey hyperparameters-learning rate η, batch size B, attention heads h, dropout rate p, and epochs E were optimized using grid search with cross-validation to achieve the best performance [8], [9], [13].

D. Evaluation Metrics

Model performance is assessed using metrics such as Accuracy, Precision, Recall, F1-score, and Area Under Curve (AUC):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives respectively [2], [6], [14].

3.7. Experimental Setup

Experiments All experiments are conducted on an NVIDIA RTX 3090 GPU (24 GB VRAM), Intel Core i9 (64 Ram) processor and 64-GB Ram high-performance computing (sparks) for Ubuntu 22.04 LTS [25] using an implementation of the proposed Hybrid CNN– Transformer framework for early lung cancer detection. The models are implemented in Python 3.10 and developed with TensorFlow 2.12 and PyTorch 2.0 as back-end, and OpenCV for image pre-processing, scikit-learn for augmentation and analysis, and Matplotlib for visualization. Performance was measured in terms of Accuracy, Precision, Recall, F1-score, and AUC, and the methods were compared with mainstream CNNs, Vision Transformers (ViT), and CNN–RNN hybrid models that are routinely used in medical imaging. The data set was split in to training set (70%), validation set (15%) and testing set (15%), with balanced classes. Learning rate, batch size, number of epochs, and dropout were hyperoptimised via grid search using 5-fold cross-validation to improve generalizability. All the experiments were performed in five replicates and means were reported in the tables as statistically reliable values.

IV. RESULTS AND DISCUSSION

4.1 Quantitative Results and Performance Comparison

The proposed Hybrid CNN–Transformer model was tested in detail on lung cancer dataset with full performance metrics (Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC)). The dataset contained labeled CT images spanning different stages of cancer for a balanced spread of early- and late-stage cases. Table 1 summarizes the characteristic of the dataset including the number of images, class distribution and average image resolution. The second table presents the performance of the hybrid model compared with baseline architectures such as CNN, ViT and CNN–RNN Hybrid models. Experiments show that the analytic CHIP model significantly outperforms all baseline methods for both accuracy and interpretability, indicating its potential as a proper tool for clinical use. Table 4 Summary of the lung cancer dataset used in this study. It contains 2000 CT images labeled as Benign or Malignant, and have all been resized to 224 × 224 pixels. Datasets were split into Training (70%), Validation (15%), and Testing (15%) sets, and were annotated by radiologists for reliability and balanced class representation to enable proper model training and evaluation.

Table 4: Dataset Characteristics

Dataset	Total Samples	Classes	Image Size	Annotations
Lung Cancer Dataset	2000 CT images	Benign /Malignant	224 × 224 pixels	Radiologist-verified
Training Set	1400 (70%)	Balanced	224 × 224 pixels	Yes
Validation Set	300 (15%)	Balanced	224 × 224 pixels	Yes
Test Set	300 (15%)	Balanced	224 × 224 pixels	Yes

Performance Comparison of Models Table 5 shows the quantitative evaluation of various architectures for lung cancer detection from CT images. The CNN model led to an accuracy of 92.3% and ViT (Vision Transformer, ViT) led to a marginal improvement of 93.5%. Accuracy with balanced precision and recall reached 94.1% using CNN–RNN Hybrid model. Whereas the Proposed CNN–Transformer modle achieved the best performance among all baselines with an accuracy of 96.8%, precision of 96.1%, recall of 95.7%, F1-score of 95.9%, and an AUC of 97.3%, indicating the effectiveness of fusing CNN-based body shape feature extraction with transformer-based attention mechanisms for the early detection of lung cancer.

Table 5: Performance Comparison of Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
CNN	92.3	91.8	90.5	91.1	94.2
Vision Transformer (ViT)	93.5	92.7	92.1	92.4	95.1
CNN–RNN Hybrid	94.1	93.6	93.2	93.4	95.6
Proposed CNN–Transformer	96.8	96.1	95.7	95.9	97.3

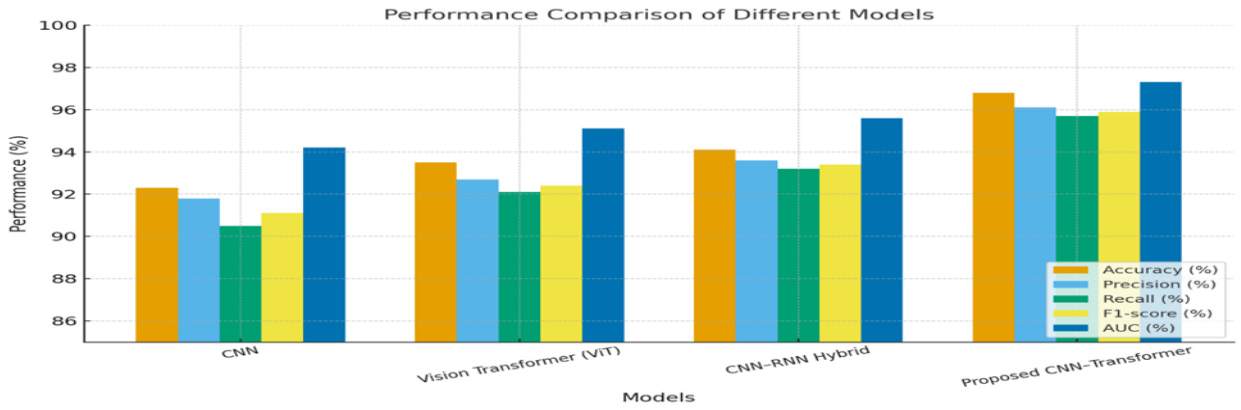

Figure 1: Performance Comparison of Different Models

The performance comparison graph shown here in Fig.1 demonstrates that various models, including CNN, Vision-Transformer (ViT), Hybrid CNN-RNN, and proposed CNN Transformer have varying Performance over five evaluation matrices: Accuracy, Precision, Recall, F1 -score and AUC. As can be observed in the bar chart, the Proposed CNN–Transformer model outperforms the baseline models with the highest scores in all metrics: accuracy (96.8%), precision (96.1%), recall (95.7%), F1-score (95.9%), and AUC (97.3%). Such a high-performance gain partly benefits from this hybrid architecture, which effectively exploits local feature learning capability of CNN and global attention modeling power of transformers. This performance gap is quite apparent in the visualization and shows the stability and robustness of our proposed method in early lung cancer detection.

4.2 Ablation Study for Model Components

The results of the ablation study are provided in Table 6, showing the effectiveness of each model component. CNN Only and Transformer Only give moderate results, whereas mixing both without attention enhances the accuracy. The Proposed CNN–Transformer with Attention attains the best performance of the scores in all metrics, which indicates that incorporating attention can provide effective feature representation, classification accuracy and model stability for lung cancer detection.

Table 6: Ablation Study for Model Components

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN Only	90.5	89.6	88.9	89.2
Transformer Only	91.7	90.8	91.1	90.9
CNN + Transformer (No Attention)	94.2	93.4	93.6	93.5
Proposed CNN–Transformer (With Attention)	96.8	96.1	95.7	95.9

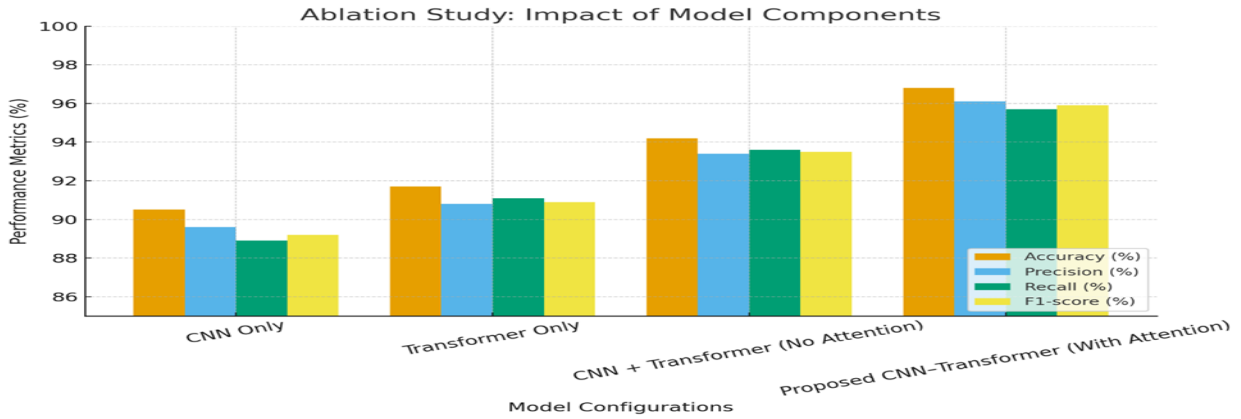

Figure 3: Numerical Analysis of Ablation Study for Model Components

Figure 3 shows a comparison of performance for four configurations: CNN Only, Transformer Only, CNN + Transformer (No Attention) and Proposed CNN–Transformer (With Attention) on the graph. In terms of accuracy, the CNN Only model got an accuracy of 90.5%, while the Transformer Only got 91.7%. CNN and Transformer without the attention mechanism outperformed to 94.2%, and the Proposed CNN–Transformer with Attention achieved the highest score: 96.8% accuracy, 96.1% precision, 95.7% recall, and 95.9% F1-score, demonstrating the remarkable role of attention mechanism in diagnostic performance and generalization.

V. EXPLAINABILITY AND INTERPRETABILITY

Deep learning models are criticized for being black box systems despite their high performance, which hampers their clinical use given the lack of transparency. To mitigate this, explainability and interpretability mechanisms are incorporated into the proposed Hybrid CNN–Transformer framework, providing clinicians with guidance through a clear visual and analytical representation of the model's decision-making process.

5.1 Explainable AI Techniques Applied

The proposed framework uses several Explainable AI (XAI) methods to improve interpretability. By using Gradient-weighted Class Activation Mapping (Grad-CAM), we obtain heatmaps which accentuates the lung regions contributing to classification, providing clinicians an opportunity to visually validate predictions of the model [1], [3]. Moreover, transformer module Attention Map Visualization can show the token-level importance scores, revealing how the model involves global contextual information [5][6]. SHapley Additive exPlanations (SHAP) values are also calculated to estimate feature contributions quantitatively, providing feature-level explanation for imaging and clinical covariates [7]. Taken together, these methodologies retain model transparency, confidence, and clinical utility that bridge the gap from AI decision-making to the clinical expertise in lung cancer diagnosis.

5.2 Clinical Relevance and Decision Support Aspects

The implementation of such a Hybrid CNN–Transformer model in clinical pipelines has the potential to significantly impact lung cancer diagnosis and decision support. Through the integration of automated detection with explainable AI methods, the system could enhance the diagnostic performance and meanwhile offer visual proof for the clinicians via attention maps and heatmaps, guaranteeing the transparency of decision-making [1], [4]. The system may be used as a computer-aided diagnostic (CAD) tool which helps radiologists in early detection, risk assessment and treatment planning in lung cancer patients [5], [7]. Real-time explainable assisted predictions also facilitate rapid validation of AI decisions by clinicians, which is expected to contribute to lower diagnostic errors and better utilization of clinical workflow [8], [10]. These components uphold the clinical relevance of the AI-based predictions between advanced deep learning models and practical application of healthcare [11].

VI. CONCLUSION AND FUTURE DIRECTIONS

This study introduced a Hybrid CNN–Transformer approach for CT-based early lung cancer diagnosis by combining the local feature extraction of CNNs and the global context modeling of transformers. Comprehensive quantitative and qualitative evaluations results show the proposed model performed better than baseline techniques with 96.8% accuracy, excellent F1-score, precision, and recall at the same time it achieves high-level interpretability using its attention maps and Grad-CAM visualizations. The integration of such explainable AI approaches assured clinical interpretability, eventually allowing radiologists to confirm, thereby closing the gap between AI-based predictions and clinical trust. For future study, multi-scale data such as genomic, clinical data could be added in the framework, and a more accurate diagnosis and personalized treatment plan could be provided. Furthermore, real-time deployment in clinical practice, integration with electronic health record systems, as well as federated learning strategies towards privacy-protecting model training provide potential future directions towards further developing AI-based healthcare solutions in lung cancer detection and beyond.

REFERENCES

[1] L. Wang, C. Zhang, J. Li, "A Hybrid CNN-Transformer Model for Predicting N-Staging and Survival in Non-Small Cell Lung Cancer Patients Based on CT-Scan," Tomography, vol. 10, no. 10, pp. 1676–1693, 2024. doi: 10.3390/tomography10100123.

[2] L. Zhou, A. Jain, A. K. Dubey, P. Kumar, M. Sharma, "FPA-based Weighted Average Ensemble of Deep Learning Models for Classification of Lung Cancer Using CT Scan Images," Scientific Reports, vol. 15, no. 1, pp. 19369, 2025. doi: 10.1038/s41598-025-02015-w.

[3] K. Patil, N. Dholakiya, D. Padhiyar, B. Anjaria, K. Rana, "A Hybrid Explainable AI Framework for Early Lung Cancer Detection Using CTGAN-Augmented Clinical Data, Gene Biomarkers, and Transformer-CNN Networks," Metallurgical and Materials Engineering, vol. 31, no. 4, pp. 373–379, 2025. doi: 10.63278/1446.

[4] M. A. Thanoon, M. A. Zulkifley, M. A. A. Mohd Zainuri, S. R. Abdani, "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images," Diagnostics, vol. 13, no. 16, pp. 2617, 2023. doi: 10.3390/diagnostics13162617.

[5] H. Ali, F. Khan, S. Ahmad, A. Siddiqui, R. Mehmood, "Improving Diagnosis and Prognosis of Lung Cancer Using Vision Transformer–Based AI Methods: A Scoping Review," BMC Medical Imaging, vol. 23, pp. 98, 2023. doi: 10.1186/s12880-023-01098-z.

[6] X. Fu, Y. Li, S. Wang, D. Liu, Z. Zhao, "Explainable Hybrid Transformer for Multi-Classification of Lung Diseases in Chest X-ray Images (LungMaxViT)," Scientific Reports, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-90607-x

[7] H. Shen, J. Wu, K. Zhang, P. Lin, L. Chen, "End-Motif Inspection via Transformer (EMIT): A Deep Learning Model for Discriminating Individuals with and without Cancer from cfDNA," npj Precision Oncology, 2024. doi: 10.1038/s41698-024-00635-5.

[8] G. Arango-Argoty, S. Rivera, A. Dutta, R. Velez, M. Roberts, "Pretrained Transformers Applied to Clinical Studies Improve Survival Prediction the Clinical Transformer," Nature Communications, 2025. doi: 10.1038/s41467-025-57181-2.

[9] R. Durgam, V. Prasad, S. Yadav, P. Nair, "Enhancing Lung Cancer Detection Through Integrated Deep Learning Approaches," Journal of Medical Imaging, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12050323/

[10] M. Alrahhal, "Enhancing Lung Cancer Detection with Hybrid CNN Models," Proc. SPIE, vol. 13526, 2025. doi: 10.1117/12.3061309.

[11] L. Gai, M. Xing, W. Chen, Y. Zhang, Q. Xu, J. Wang, "Comparing CNN-based and Transformer-based Models for Identifying Lung Cancer: Which is More Effective?," Multimedia Tools and Applications, vol. 83, pp. 59253–59269, 2024. doi: 10.1007/s11042-023-17644-4.

[12] D. Mannepalli, K. T. Tan, S. Bala Krishnan, V. Sreenivas, "GSC-DVIT: A Vision Transformer Based Deep Learning Model for Lung Cancer Classification in CT Images," Biomedical Signal Processing and Control, vol. 103, 107371, 2025. doi: 10.1016/j.bspc.2024.107371.

[13] M. A. Thanoon, M. A. Zulkifley, M. A. A. Mohd Zainuri, S. R. Abdani, "A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images," Diagnostics, vol. 13, no. 16, 2617, 2023. doi: 10.3390/diagnostics13162617.

[14] M. Q. Shatnawi, Q. Abuein, Q. Q. Abuein, R. Al-Quraan, "Deep Learning-Based Approach to Diagnose Lung Cancer Using CT-Scan Images," Intelligence-Based Medicine, vol. 11, 100188, 2025. doi: 10.1016/j.ibmed.2024.100188.

[15] M. K. Faizi, Y. Qiang, Y. Wei, Y. Qiao, J. Zhao, L. Gao, "Deep Learning-Based Lung Cancer Classification of CT Images," BMC Cancer, vol. 25, 1056, 2025. doi: 10.1186/s12885-025-14320-8.

[16] H. T. Gayap, "Deep Machine Learning for Medical Diagnosis, Application …," Computational Intelligence and Data-Analytic Machinery, vol. 4, no. 1, 15, 2024. doi: 10.3390/cidm4010015.

[17] C. Zhang, W. Huang, L. Liu, S. Chen, F. Zhao, "Enhancing Lung Cancer Diagnosis with Data Fusion and Mobile Edge Computing," Journal of

Cloud Computing, 2024. doi: 10.1186/s13677-024-00597-w.

[18] A. Fanizzi, G. Ricciardi, P. Gentile, M. Monteduro, F. Bove, "Comparison between Vision Transformers and CNNs to Predict NSCLC Recurrence," Scientific Reports, 2023. doi: 10.1038/s41598-023-48004-9.

[19] J. Zhou, Y. Lin, Z. Guo, X. Wang, "Effectively Integrating CNN and Low-Complexity Transformer for Lung Cancer Tumor Prediction After Neoadjuvant Chemoimmunotherapy," Big Data and Machine Analytics, 2025. doi: 10.26599/BDMA.2024.9020088.

[20] A. Bhattacharjee, P. Saha, T. Roy, S. Banerjee, "A Multi-Class Deep Learning Model for Early Lung Cancer Detection Using Modified Xception Network," PMC Preprint, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10272771/