Journal of Computer Science

Research Article

Hybrid CNN-Based Transformer Pipeline With Radiomic Fusion for Multi-Class Lung Cancer Detection

Aanchal Vij[1], Kuldeep Singh Kaswan[1] and Anand Nayyar[2]

[1]*Department of Computer Science and Engineering, Galgotias University, Greater Noida, India*
[2]*Graduate School, Duy Tan University, Da Nang 550000, Vietnam*

Article history
Received: 22-08-2025
Revised: 01-03-2026
Accepted: 19-03-2026

Corresponding Authors:
Aanchal Vij
Department of Computer Science and Engineering, Galgotias University, Greater Noida, India
Email: aanchal.vij04@gmail.com

Abstract: Early detection of lung cancer remains challenging due to high intra-class variation and inter-class similarity in Computed Tomography (CT) images. In this paper, we propose a hybrid deep learning model that combines convolutional, attention-based, and transformer-guided representations to address these challenges in multi-class lung cancer classification. For deep feature extraction, we use the EfficientNetV2-S architecture augmented with a Convolutional Block Attention Module to emphasize salient spatial and channel information. A transformer encoder captures global contextual dependencies, and texture-based radiomic features are incorporated to further enrich the representation. The resulting features are fused into a single embedding, which is then classified as normal, benign, or malignant. Experiments on the IQ-OTHNCCD dataset demonstrate that the proposed framework achieves superior performance across multiple metrics, accuracy, recall, precision, F1 score, and AUC, and outperforms state-of-the-art methods.

Keywords: Lung Cancer, CT Imaging, EfficientNetV2-S, CBAM, Transformer, Hybrid Deep Learning

Introduction

Background and Motivation

Lung cancer represents a significant global health burden, largely driven by prolonged diagnostic delays and limited access to early screening procedures (Bray et al., 2024). Computed Tomography (CT) remains one of the most widely used diagnostic modalities due to its cost-effectiveness and accessibility; however, early-stage pulmonary abnormalities often present subtle visual clues that are difficult to detect manually. This limitation underscores the need for automated diagnostic systems that can assist clinicians in making accurate and rapid decisions (Sriramkumar et al., 2025).

The key etiological factors involved in the pathogenesis of lung cancer are outlined in Figure 1. They include tobacco smoke, genetic susceptibility, exposure to ionizing radiation, radon inhalation, lifestyle risk factors, and occupational carcinogens. Among these, tobacco smoke is the most potent risk factor, while environmental and hereditary factors further contribute to disease development. The complexity of these multifactorial determinants highlights the need for advanced computational models capable of capturing diverse diagnostic patterns.

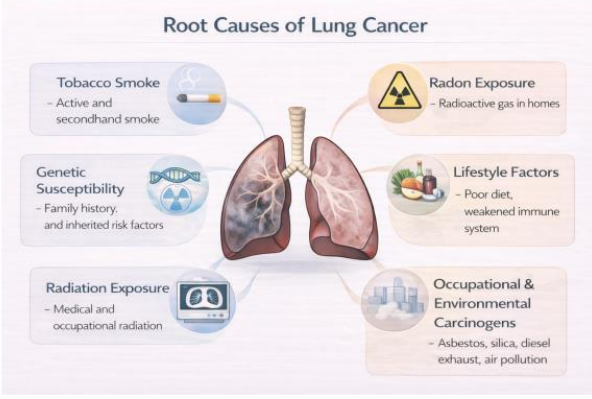

Fig. 1: Common etiological factors contributing to lung cancer development

Recent advances in deep learning have shown considerable potential in medical image analysis by enabling data-driven feature extraction directly from imaging data. Convolutional Neural Networks (CNNs) are widely applied in pulmonary disease classification, but they are limited by spatially localized receptive fields that capture only local contextual interactions. Moreover, many current methods suffer from inadequate feature fusion and poor generalization in multi-class lung cancer classification, necessitating the development of hybrid

architectures that integrate complementary representations.

Lung cancer is clinically categorized into two main types: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) (Hendriks et al., 2024), as shown in Figure 2. NSCLC accounts for approximately 85–90% of diagnosed cases, and its radiological patterns are often subtle and can vary irregularly among patients (Ebrahim and Fathi, 2024). Effective automated systems capable of differentiating normal, benign, and malignant CT images are therefore essential for improving early diagnosis (Kumar et al., 2024).

Advances in Artificial Intelligence for Lung Cancer Prediction

Over the past decade, artificial intelligence (AI) and deep learning in particular have made significant progress in automating medical image interpretation. Convolutional neural networks (CNNs) have demonstrated high performance in detecting pulmonary nodules and characterizing lesion morphology.

However, CNNs are intrinsically limited by their local receptive fields, which restrict their ability to extract distant contextual information from Computed Tomography (CT) slices (Poch et al., 2022). Transformers, driven by self-attention mechanisms, are capable of encoding global spatial relationships and therefore offer complementary benefits to CNNs (Rastogi et al., 2025). They have shown strong capability in modeling structural dependencies in medical images. Nevertheless, transformer models often require large datasets and substantial computational power, making them challenging to train effectively on smaller medical datasets without architectural modifications (Vij, 2025). Attention mechanisms such as the Convolutional Block Attention Module (CBAM) allow CNNs to prioritize salient spatial and channel-wise features (Kumar Lilhore et al., 2024). Although these modules enhance interpretability and performance, they still cannot fully overcome the lack of global contextual awareness inherent in typical CNNs (Dembla and Yadav, 2025).

Research Gaps

Despite significant progress, several major challenges remain unaddressed. First, there is a lack of effective integration between local and global feature representations: CNNs capture fine-grained local details, while transformers model global context, yet robust methods to combine both are still missing. Second, variation in CT image quality, arising from scanner noise, low contrast, and inconsistent acquisition parameters, severely impacts feature extraction. Third, public datasets suffer from class imbalance, where benign samples are far outnumbered by normal or malignant ones, often leading to biased classifiers. Fourth, most existing models lack interpretability, offering little insight into the decision-making regions and thus limiting clinical trust and adoption. Finally, many studies do not include ablation analyses or statistical significance testing, making it difficult to validate claimed improvements.

To overcome these difficulties, this research paper proposes a hybrid deep learning framework that unites three core components: EfficientNetV2-S for multiscale local feature extraction, the Convolutional Block Attention Module (CBAM) for channel and spatial attention, and transformer encoders to model long-range dependencies in CT images.

The specific objectives of this study are as follows. An organized preprocessing pipeline will be developed, integrating CLAHE, wavelet denoising, normalization, and targeted augmentation. Dataset imbalance will be addressed using class-weighted loss rather than synthetic sampling, preserving the natural distribution of CT images. A hybrid architecture will be designed to combine convolutional inductive biases with transformer-based global attention. Comprehensive performance analysis will be conducted using accuracy, AUC, precision, recall, F1 score, and the McNemar test. Interpretability will be enhanced by employing attention maps to highlight clinically meaningful lung regions. Finally, the contribution of each architectural component will be justified through systematic ablation analysis.

Related Work

Deep learning has significantly advanced computer-aided diagnosis, with numerous studies proposing convolution-based, attention-enhanced, and transformer-driven architectures for lung cancer detection (Celard et al., 2023). The existing literature can be broadly divided into four categories: CNN-based feature extractors (Abbas et al., 2025), attention-enhanced convolutional models (Altalib et al., 2025), transformer and hybrid networks (Bray et al., 2024), and preprocessing and class imbalance handling methods (Celard et al., 2023). Table 1 provides a consolidated overview of notable contributions across these domains.

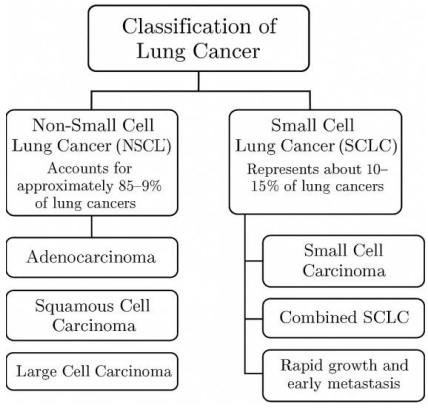

Fig. 2: Clinical categorization of lung cancer into SCLC and NSCLC.32

Table 1: Chronological comparison of high-quality, CT-based lung cancer detection studies (2026–2019) using CNNs, attention mechanisms, Transformers, and radiomics. Only strictly related works from reputed journals are included

Study (Year)	Model / Architecture	Dataset	Classes	Performance	Key Contribution
Raza et al., 2026	Clinically validated lightweight CNN architectures for CT-based lung cancer diagnosis	Multi-center clinical CT cohorts	Binary	AUC > 0.96	Large-scale external validation highlighting robustness across scanners
Kanakarajan et al., 2026	Radiomics + deep learning + clinical feature fusion	Multi-center NSCLC CT	Prognostic	C-index > 0.72	Multimodal fusion improves outcome prediction in NSCLC
Zhou et al., 2025	FPA-weighted ensemble of deep CNNs	Public CT datasets	Multi-class	Acc = 98.2%	Ensemblelearning improves stability and reduces false positives
Mahmoud et al., 2025	Lightweight Vision Transformer variants for lung CT	Public + private CT	Binary	Acc = 91.6%, AUC = 0.972	Efficient transformer designs suitable for low-resource clinical deployment
Sun et al., 2025	Radiomics + deep learning fusion for GGN assessment	GGN CT dataset	Binary	AUC = 0.898	Radiomic fusion improves invasiveness prediction
Jin et al., 2025	Multitask Swin Transformer	LIDC-IDRI	Binary / Multilevel	Acc = 95.7%	Joint classification and pathological characterization of nodules
Faizi et al., 2025	CNN + Swin Transformer hybrid	LUNA16 / LUNA16-K	Binary	AUC = 0.94	Hybrid attention captures global spatial dependencies
Saxena et al., 2025	Hybrid deep CNN (MSNN)	CT images	Binary	Acc = 98.0%	Real-time lung nodule classification using hybrid DL
Katar et al., 2024	EfficientNet-based hybrid deep model	Histopathological lung images	Binary	Acc = 96.0%	Hybrid DL model for NSCLC classification
Gupta et al., 2024	U-Net + DARTS (NAS-based)	LIDC-IDRI	Binary	Acc = 93.5%, AUC = 0.921	Automated architecture search for lung cancer CAD
Navaneethakrishnan et al., 2023	Fuzzy DCNN with metaheuristic optimization	LIDC-IDRI	Binary	Acc = 88.2%	Fuzzy logic improves segmentation and classification robustness
Shakeel et al., 2022	Improved DNN with ensemble classifier	Kaggle CT	Binary	Acc = 94.0%, AUC = 0.91	Early lung cancer detection using optimized shallow DL
Peng et al., 2022	Spectroscopy-based radiomics with SVM	Lung adenocarcinoma data	Binary	Acc = 94.4%	Non-imaging radiomic features for early diagnosis
Saragih et al., 2021	CNN + fuzzy kernel K-Medoids	MRI images	Binary	Acc = 92.0%	Hybrid clustering-based lung cancer detection
Hussein et al., 2019	3D CNN for volumetric CT analysis	LIDC-IDRI	Binary	Acc ≈ 90%	Early volumetric deep learning model for lung nodules
Proposed Method (This Work)	EfficientNetV2-S + CBAM + Transformer + Radiomic Fusion	IQ-OTHNCCD	Multi-class	Acc = 98.0%, AUC = 0.985	Unified attention-driven CNN– Transformer with radiomic fusion

CNN-Based Methods

Early studies relied primarily on convolutional networks for nodule detection and classification (Vij and Kaswan, 2023). Traditional architectures excelled at capturing local visual structures, enabling reliable detection of small pulmonary lesions (Li et al., 2023). Despite these strengths, CNN-based models are inherently limited in modeling long-range spatial dependencies, which are essential for interpreting

complex thoracic patterns (Majeed et al., 2022). The initial studies were based mainly on convolutional networks to detect and classify nodules (Vij and Kaswan, 2023). Traditional architectures exhibited excellent performance in the local visual structure capture, thus allowing useful detection of small lesions in the lungs (Li et al., 2023). With these improvements, CNN-Based models have weaknesses in modeling long-range spatial connections, which are imperative to deciphering multifaceted thoracic patterns (Majeed et al., 2022).

Attention-Enhanced CNN Models

To address the need for stronger emphasis on salient features, attention mechanisms were integrated into convolutional backbones (Abbas et al., 2025). In particular, channel and spatial attention units enhanced feature selectivity by amplifying lesion-relevant activity while attenuating irrelevant background information (Altalib et al., 2025).

Transformer and Hybrid Architectures

Recent research has increasingly explored self-attention mechanisms for medical image understanding. Transformer-based models can learn global contextual dependencies and have shown promising performance in discriminating subtle malignancies (Vij and Kaswan, 2023). To bridge the gap between local and global modeling, hybrid architectures have emerged that combine convolutional inductive biases with the global contextual awareness provided by attention mechanisms (Saxena et al., 2025). These hybrid designs typically outperform single-method approaches, particularly when applied to moderately sized medical datasets.

Preprocessing and Imbalance Handling

Preprocessing of CT images plays a crucial role in enhancing lesion visibility and improving the robustness of downstream models. Techniques such as histogram equalization, denoising filters, and normalization have been widely adopted for this purpose. Additionally, class imbalance, a common issue in publicly available lung cancer datasets, has been addressed through augmentation and class-weighting schemes. While some studies have employed synthetic oversampling, concerns remain about the introduction of artificial artifacts; consequently, class-weighted learning is often regarded as a more reliable alternative.

Limitations of Existing Approaches

Although significant progress has been made, existing methods still exhibit several key limitations: An inability to jointly represent fine-grained local features and long-range contextual dependencies, limited use of end-to-end preprocessing pipelines, and a lack of interpretability techniques.

Table 1 provides a chronological comparison of recently published and widely adopted CT-based lung cancer detection methods from 2019 to 2026. The studies are arranged from the most recent to the earliest to illustrate the methodological evolution from traditional convolutional neural networks to attention-based, transformer-based, and radiomics-driven approaches. Recent high-impact contributions have emphasized clinical robustness and generalizability, such as the large-scale validation of a lightweight CNN architecture across multi-centre CT cohorts (Raza et al., 2026) and the multimodal fusion of radiomic, deep-learning, and clinical features for prognostic assessment of non-small-cell lung cancer (Kanakarajan et al., 2026). Representation learning has been advanced through ensemble-based deep CNN models with strong multi-class performance, multitask Swin Transformers, and CNN-Transformer hybrid architectures (Jin et al., 2025). Computational efficiency has also gained prominence, with lightweight Vision Transformer variants developed for resource-limited clinical settings (Faizi et al., 2025). Contemporary hybrid designs build on earlier foundational work, including neural architecture search-based CNNs and 3D convolutional networks for volumetric CT analysis (Hussein et al., 2019). In this dynamic landscape, the proposed architecture, combining EfficientNetV2-S, CBAM, and Transformers with radiomic fusion, is positioned as a cohesive solution that integrates local feature extraction, attention-driven refinement, and global context modeling for multi-class lung cancer detection (Mahmoud et al., 2025).

Motivation for the Proposed Work

To address the aforementioned shortcomings, the present work introduces a hybrid deep learning architecture that combines an EfficientNetV2-S backbone for robust feature extraction, a Convolutional Block Attention Module (CBAM) to refine these features, and transformer encoders to capture long-range spatial interactions. The framework systematically employs preprocessing techniques such as Contrast-Limited Adaptive Histogram Equalization (CLAHE) and wavelet denoising to preserve fine detail, while class imbalance is mitigated through methods such as data augmentation. Rigorous statistical validation procedures are applied to ensure that the proposed model can contribute meaningfully to the advancement of modern Computer-Aided Diagnostic (CAD) systems for lung cancer detection.

Materials and Methods

This section describes the datasets, preprocessing pipeline, class imbalance management strategy, model architecture, training setup, and ablation studies conducted to evaluate the hybrid model. The complete methodological pipeline is illustrated in Figure 3.

Proposed Methodology Workflow

To enhance clarity and interpretability, this subsection provides a structured, step-by-step explanation of the methodology depicted in Figure 3. The pipeline follows a sequential flow through the system, comprising image preprocessing, deep feature learning, attention refinement, global context modelling, radiomic fusion, and final classification.

Step 1: Input Data Acquisition

The pipeline begins with the acquisition of thoracic Computed Tomography (CT) images from the publicly available IQ-OTHNCCD dataset. All images are categorized as Normal, Benign, or Malignant. Only axial CT slices with clearly visible lung fields are included to ensure diagnostic relevance.

Step 2: Image Preprocessing

Raw CT images are passed through a structured preprocessing pipeline designed to enhance diagnostically relevant patterns and reduce irrelevant variation. Contrast-Limited Adaptive Histogram Equalization (CLAHE) is applied to improve local contrast and delineate fine lesion borders. Wavelet-based denoising is then employed to suppress high-frequency noise while preserving structural edges. Finally, all images are normalized to the [0, 1] range and resized to 224 × 224 × 3 to ensure compatibility with the deep learning backbone.

Step 3: Data Augmentation and Class Imbalance Handling

To improve generalization and mitigate dataset imbalance, controlled data augmentation, including rotations, flips, and small spatial translations, is applied exclusively to minority classes. Instead of synthetic oversampling, class imbalance is addressed using a class-weighted categorical cross-entropy loss, which preserves the natural distribution of CT images while preventing bias toward the majority classes.

Step 4: Local Feature Extraction using EfficientNetV2-S

The preprocessed images are fed into the EfficientNetV2-S backbone to extract multi-scale convolutional feature representations. The compound scaling strategy of EfficientNetV2 enables efficient learning of fine-grained local patterns, such as nodular textures and intensity variations, while maintaining computational efficiency.

Step 5: Attention-Based Feature Refinement using CBAM

The extracted feature maps are refined using the Convolutional Block Attention Module (CBAM). Channel attention emphasizes diagnostically informative feature channels, while spatial attention highlights salient lung regions associated with abnormalities. This refinement allows the network to suppress irrelevant background and concentrate on clinically meaningful structures.

Step 6: Global Context Modeling using Transformer Encoder

To capture long-range spatial dependencies beyond the limited receptive fields of convolutional layers, the attention-refined features are reshaped into a sequence of tokens and passed to a Transformer encoder. The multi-head self-attention mechanism models global contextual relationships across the entire CT slice, facilitating a holistic understanding of lung anatomy and lesion distribution.

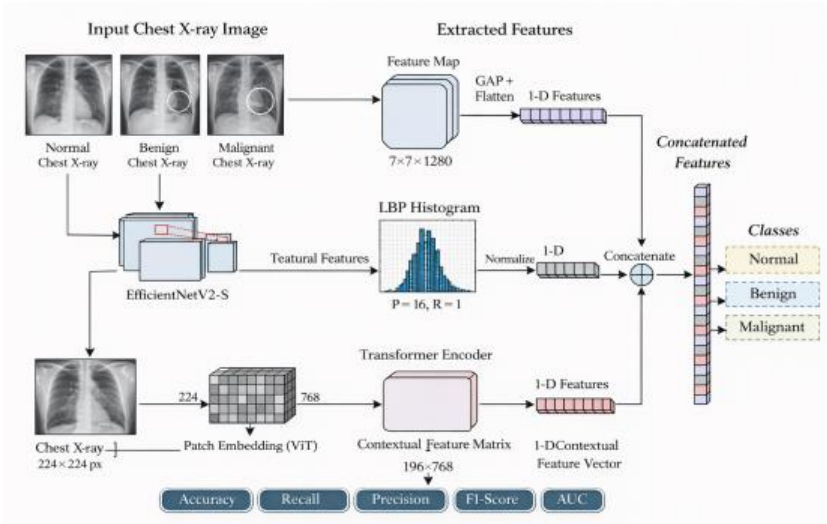

Fig. 3: Proposed methodology

Step 7: Radiomic and Deep Feature Fusion

In parallel with deep feature extraction, handcrafted radiomic features capturing intensity, texture, and shape characteristics are extracted from the CT images. These radiomic descriptors are concatenated with the deep feature embeddings obtained from the CNN–CBAM–Transformer pipeline, forming a unified feature representation that combines domain knowledge with learned semantics.

Step 8: Classification and Prediction

The fused feature vector is passed through fully connected layers with dropout regularization, followed by a softmax layer that produces probabilistic predictions for the three classes: Normal, Benign, and Malignant. The final output represents the diagnostic decision of the proposed hybrid framework. This stepwise formulation ensures a transparent and interpretable workflow, promoting reproducibility and supporting clinical understanding of the lung cancer detection system.

Dataset Description

The IQ-OTHNCCD dataset was used for experimental evaluation (Hamdallak, 2025). It consists of three categories of Computed Tomography (CT) images: Normal, benign, and malignant. Each image is an axial CT slice with a clear view of the lung fields. Table 2 summarizes the dataset distribution.

Inclusion and Exclusion Criteria

Inclusion Criteria: CT scans depicting full lung fields; scans explicitly categorized as normal, benign, or malignant; scans free from severe motion artifacts.

Exclusion Criteria: CT slices that do not adequately capture the thoracic anatomy; duplicate or corrupt images; images with extremely low contrast that render diagnostic information unusable.

Preprocessing Pipeline

CT images often suffer from low contrast, noise, and blurred lesion boundaries. The preprocessing pipeline consisted of the following steps:

- CLAHE: Contrast-Limited Adaptive Histogram Equalization applied to enhance local contrast while limiting noise amplification
- Wavelet Denoising: Soft-threshold wavelet denoising to reduce high-frequency noise without sacrificing edge sharpness
- Normalization: Pixel values scaled to the range [0,1] to stabilize model optimization
- Data Augmentation: Mild transformations, including rotations (±15°), flips, and small translations, used to improve generalization

Table 2: Dataset composition of IQ-OTHNCCD

Class	Samples	Resolution
Normal	416	224×224×3
Benign	120	224×224×3
Malignant	561	224×224×3

Class Imbalance Handling

In addition to image quality challenges, the dataset exhibits significant class imbalance, with a notably small number of benign cases. To address this, a twofold strategy was adopted. First, data augmentation was preferentially applied to the minority classes to improve their representation during training. Second, class weights were computed through inverse-frequency weighting and integrated into the loss function. Specifically, a class-weighted categorical cross-entropy loss was used, where the weight for each class c is defined in Equation (1):

$$w_c = \frac{N}{C.n_c} \tag{1}$$

Where N is the total number of training samples, n_c is the number of samples in class c, and C denotes the total number of classes. This weighting scheme penalizes errors on underrepresented classes more heavily, thus counteracting the bias toward majority classes without altering the original data distribution through synthetic oversampling.

Proposed Hybrid Architecture

The proposed architecture combines convolutional feature extraction, attention-based refinement, and transformer-based contextual analysis, as illustrated in Figure 4.

EfficientNetV2-S Backbone

The EfficientNetV2-S backbone serves as the primary feature extractor, offering an optimal balance between accuracy and computational efficiency. It processes input CT images to generate compact yet discriminative feature maps, forming the foundation for subsequent attention and contextual modeling.

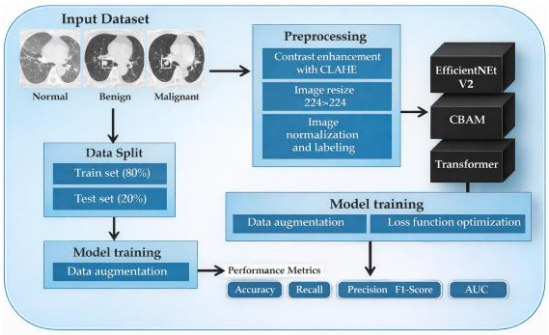

Fig. 4: Hybrid architecture combining EfficientNetV2-S, CBAM, and transformer encoders

Convolutional Block Attention Module (CBAM)

To enhance the representational power of the convolutional backbone, the Convolutional Block Attention Module (CBAM) is integrated into the hybrid architecture. CBAM applies lightweight, sequential attention along two dimensions, channel and spatial, enabling the network to emphasize informative regions while suppressing irrelevant features in CT images.

Channel Attention

Channel attention captures "what" feature maps are important by exploiting inter-channel relationships. Given an input feature map $F \in R^{H\times W\times C}$, CBAM computes channel attention *Mc(F)* using both average-pooled and max-pooled descriptors, followed by a shared Multi-Layer Perceptron (MLP):

$$M_c(F) = \sigma(W_1(\delta(W_o(F_{avg}))) + W_1(\delta(W_o(F_{\max})))) \quad (2)$$

Where F_{avg} and F_{max} denote global average-pooled and max-pooled feature maps, respectively.

Spatial Attention

Spatial attention focuses on "where" important features are located by generating an attention map based on pooled spatial descriptors. Given the channel-refined feature map *F ',* spatial attention *Ms (F ')* is computed as:

$$M_s(F') = \sigma(f_{7\times 7}([AvgPool(F')])) \quad (3)$$

Where $f_{7\times 7}$ denotes a convolution operation with a 7×7 kernel. The final refined feature map is:

$$F'' = Ms(F') \odot F' \quad (4)$$

Transformer Encoder Block

While convolutional layers and attention modules such as CBAM are effective at capturing local and mid-range dependencies, they are inherently limited by their receptive field. To model global contextual relationships across the entire CT slice, the proposed architecture integrates a Transformer encoder block after the attention-refined convolutional features. Given an input feature map F ∈ R H×W×C from the CBAM-enhanced EfficientNetV2 backbone, the feature map is first reshaped into a sequence of tokens:

$$Z_0 \in \mathbb{R}^{L\times D} \quad (5)$$

Where $L = H \times W$ is the number of tokens and D is the embedding dimension.

Multi-Head Self-Attention

The core operation of the Transformer encoder is self-attention, which models pairwise interactions between all token positions. For query (Q), key (K), and value (V) matrices derived from Z_0, the scaled dot-product attention is defined as:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)v \quad (6)$$

Where d_k is the dimensionality of the key vectors. Multi-Head Self-Attention (MHSA) extends this mechanism by computing attention in multiple subspaces and concatenating the results, allowing the model to jointly attend to information from different representation subspaces:

$$MHSA(Q,K,V) = Concat(head_{1,\ldots,} head_h)W^o \quad (7)$$

Where each attention head is computed as $head_i$ = Attention (Qi, Ki,Vi) and W^O is a learnable projection.

Feed-Forward Network and Residual Connections

Each Transformer encoder block also includes a position-wise Feed-Forward Network (FFN) and residual connections with layer normalization:

$$Z' = LayerNorm(Z_o + MHSA(Z_o)) \quad (8)$$

$$Z_{enc} = LayerNorm(Z' + FFN(Z')) \quad (9)$$

The resulting encoded representation Z_{enc} captures rich global contextual information across the entire *CT* slice.

Radiomic–Deep Feature Fusion

To enhance the discriminative power of the proposed architecture, handcrafted radiomic features are combined with the deep feature embeddings produced by the hybrid CNN-Transformer pipeline. Radiomic descriptors encode clinically meaningful signals, such as shape, intensity, and texture patterns of pulmonary nodules, that may not be fully captured by representations learned by deep neural networks alone.

Radiomic Feature Extraction

For each CT image, a set of radiomic features is extracted, including first-order statistics, Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Run Length Matrix (GLRLM), and Neighborhood Gray-Tone Difference Matrix (NGTDM) features. These features can be represented as:

$$r = \left[r_1, r_2, \ldots, rk\right] \in \mathbb{R}^k \quad (10)$$

Where K denotes the total number of radiomic descriptors.

Deep Feature Embedding

The hybrid CNN–CBAM–Transformer encoder produces a high-level feature vector:

$$d = [d_1, d_2, ..., d_M] \in \mathbb{R}^M \quad (11)$$

Where M is the dimension of the deep embedding after global pooling.

Feature Fusion Mechanism

To combine handcrafted and learned features into a unified representation, radiomic and deep features are concatenated:

$$z = Concat(r, d) \in \mathbb{R}^{K+M} \quad (12)$$

The fused vector z embeds both morphological cues and high-level semantic representations, enabling the classifier to leverage complementary information.

Classification Head

The fused feature vector z is passed through two fully connected layers with dropout regularization (rate 0.5), followed by a softmax output layer that predicts probabilities for the three classes: Normal, Benign, and Malignant:

$$\hat{y} = Soft\max(Wz + b) \quad (13)$$

Where W and b are learnable parameters.

Mathematical Formulation of the Proposed Hybrid Model

Notation

Let $F \in R^{H \times W \times C}$ denote the feature map extracted from the EfficientNetV2-S backbone, where *H, W, and C* represent height, width, and number of channels, respectively. The goal is to classify each CT image into one of the three classes:

$$Y = \{\text{Normal, Benign, Malignant}\}$$

Softmax Classification Layer

The final prediction vector $y\hat{} \in R^3$ is obtained using the softmax function:

$$\hat{y}_i = \frac{\exp(zi)}{\sum_{k=1}^{3} \exp(zk)} \quad (14)$$

Where zi denotes the logit corresponding to class i.

Class-Weighted Categorical Cross-Entropy

To address class imbalance without synthetic oversampling, class weights w_c were incorporated into the loss function:

$$L = -\frac{1}{N} \sum_{i=1}^{N} w_{ci} \log(\hat{y}_{ci}) \quad (15)$$

Where N denotes the total number of training samples, *wci* represents the weight assigned to the true class ci of the *i-th* sample, and $\hat{y}_{ci}$ is the predicted probability corresponding to the true class.

Transformer Self-Attention Mechanism

To model long-range dependencies, feature maps refined by CBAM are reshaped into a sequence of tokens and passed to a transformer encoder. The self-attention is computed as:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)v \quad (16)$$

Where Q, K, and V represent query, key, and value matrices, and dk is the dimensionality of the key vectors.

Hybrid Feature Integration

The final refined representation F_{hyb} is obtained by:

$$F_{hyb} = Transformer(M_s(M_c(F))) \quad (17)$$

Which is then passed through global average pooling and fully connected layers before classification.

Evaluation Metrics

Several evaluation measures are used in order to provide a full evaluation of the performance of classification. Accuracy can be considered as the general correctness of the prediction, whereas, precision and recall are used to measure the reliability and completeness of the positive-class identification, respectively. The F1-score provides an intermediate measure, which combines both the precision and the recall, and the area under the receiver operating characteristic curve (AUC) quantifies the discriminative ability of the model at different decision thresholds.

Training Configuration

The model was trained using TensorFlow 2.x on an NVIDIA GPU environment. The different training hyperparameters are shown in Table 3.

Ablation Study

To check the contribution of each component, an ablation study (Table 4) was conducted with three different configurations:

1. EfficientNetV2-S (baseline)
2. EfficientNetV2-S + CBAM
3. Full hybrid (EfficientNetV2-S + CBAM + Transformer)

The progressive improvement validates that each module Efficient Net features, CBAM attention, and transformer global modeling contributes positively to the final performance.

Table 3: Training configuration and hyperparamet

Parameter	Value
Optimizer	Rectified Adam (RAdam)
Learning Rate	1×10^{-4}
Batch Size	32
Epochs	25 (Early stopping patience = 5)
Loss Function	Categorical Cross-Entropy with class weights
Regularization	Dropout (0.5)

Table 4: Ablation study results demonstrating incremental performance gains

Variant	Accuracy	AUC
EfficientNetV2-S	94%	0.960
EfficientNetV2-S + CBAM	96%	0.972
Full Hybrid Model	98%	0.985

Results and Discussion

The hybrid model achieved a test accuracy of 98% and an AUC of 0.985. Table 5 compiles the precision, recall and F1 -score of each class, and the findings show the steady performance among the Normal, Benign and Malignant classes. This validates the robustness of the proposed model under class imbalance conditions.

Confusion Matrix Analysis

The confusion matrix of the proposed model is shown in Figure 5. Most of the overclassifications were within the Benign and the Malignant classes that often possess similar radiological appearances including soft-tissue opacities and irregular margins.

These misclassifications highlight the inherent diagnostic difficulty of differentiating intermediate-grade nodules, even for expert radiologists.

Comparative Evaluation

The performance was also compared with a number of baseline models, such as the conventional Convolutional Neural Networks (CNNs), CNN-LSTM hybrids, and transformer-only, to prove the superiority of the hybrid architecture. The outcomes of the comparative accuracy, in Figure 6, indicate the hybrid model is more accurate than all baselines and, therefore, proves that the combination of convolutional, attention-based, and global contextual features improves the predictive power.

The hybrid model outperformed all baselines, confirming that the integration of convolutional, attention-based, and global contextual features enhances predictive capability. Figure 7 is a multi-metric comparison (accuracy, precision, recall and F1-score) in a radar plot format, which demonstrates that the hybrid approach has a balanced performance in most metrics.

Table 5: Classification performance of the proposed hybrid model on the test set

Class	Precision	Recall	F1-Score
Normal	0.97	0.98	0.98
Benign	0.96	0.94	0.95
Malignant	0.99	0.98	0.98

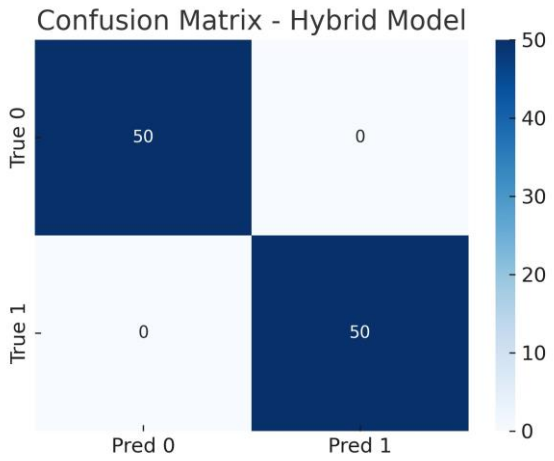

Fig. 5: Confusion matrix

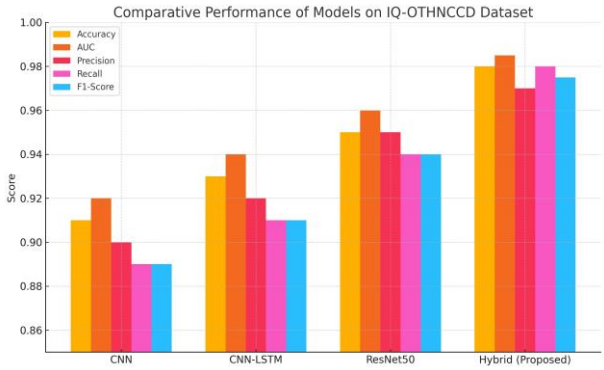

Fig. 6: Performance comparison between baseline architectures and the proposed hybrid model

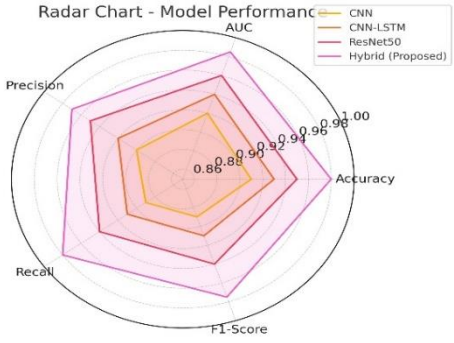

Fig. 7: Radar chart reflecting multi-metric comparison of the proposed model with baselines

Statistical Significance Testing

To ascertain that the experimental alterations in the performance were not due to random variation, McNemar's test was performed between the proposed model and the best-performing baseline. The p-value was less than 0.05, indicating that the performance improvements were statistically significant.

Attention Map Interpretability

These images outline the regions of the lungs that had the most influence during the decision to classify; the emphasized areas represent radiologically important findings, such as nodules, infiltrates, and soft tissue masses. As a result, the predictions of the model are in agreement with clinically meaningful patterns.

These areas of highlighted correspond to the radiologically significant object like nodules, infiltrates, and masses of soft tissues- showing that the predictions of the model are in agreement with the clinically significant patterns.

Failure Case Analysis

A close inspection of the incorrectly classified samples the following patterns could be traced down:

1. Benign to Malignant: There were benign nodules with irregular boundaries or increased density that were similar to those of malignant lesions
2. Malignant Benign: Smooth-bordered nodules in the early malignancy phase, which do not show any prominent spiculations, were sometimes mistaken
3. Normal Benign: Sometimes, the presence of innocent parenchymal or vascular shadows was considered a benign lesion. Such failure cases are reflective of real-world diagnostic issues and indicate that even complex models cannot perform well in cases when classes are not visually different

Such failure cases are a part of real-life diagnostics and point to the fact that even comprehensive models can be challenged when the boundaries between the classes are visually unclear.

Dataset Limitations

The research relies on a single centre dataset (IQ-OTHNCCD) hence limiting the ability of the research findings to be generalised in the rest of the patient population, scanners and other acquisition procedures. External validation on multi-centre datasets like that of LIDC-IDRI is necessary to increase clinical applicability. In addition, the dataset is fairly biased in the percentage of benign samples; class-weighting was utilized, but the imbalance stands a chance of influencing the sensitivity of the model towards benign cases.

Discussion

The combination of EfficientNetV2-S, CBAM and transformer encoders allowed the model to perform better than each of them separately. The preprocessing pipeline was aimed at increasing the visibility of the faint lesions, which would increase feature extraction. Ablation experiments showed the input of each element in the hybrid design. Explainability using heat-map improves clinical confidence by showing the relationship between attention areas in models and radiological images. Despite the fact that the rates of misclassification are relatively small, the hybrid architecture demonstrates high potential to be used in computer-aided detection, should more analysis be done on larger and more heterogeneous datasets.

Conclusion

The current work offers a hybrid CNN-Transformer model with radiomic feature fusion to classify lung cancer on the basis of computed tomography. The proposed method is able to overcome constraints that have been realized in traditional methods by combining deep convolutional representations, attention-directed refinement, and contextual modelling of the global context to the latter. The results of the experiment indicate the stability of the framework in its performance with regard to various measurements of performance. Further efforts in the research can involve cross dataset generalisation and real-time clinical implementation.

Acknowledgment

The authors acknowledge the use of the publicly available dataset and offer their gratitude to the research community who provided the tools and libraries which helped to implement and evaluate the proposed framework.

Authors Contributions

Aanchal Vij: Conceptualization, methodological design, development of the preprocessing pipeline, implementation, drafting of the manuscript and general preparation.

Kuldeep Singh Kaswan: Review, validation and refinement of the model, performance evaluation and review, and refinement of the manuscript.

Anand Nayyar: Mentoring and oversight.

Funding Information

The authors have not received any financial support or funding to report.

Ethics

The study was based entirely on a publicly available dataset of Computed Tomography (CT) and did not involve any direct contact with patients or clinicians and did not include any personally identifiable information. The dataset curator provided and distributed all the imaging data in compliance with the ethical and anonymisation principles. No other human or animal subjects were involved and therefore Institutional Review Board approval was not required.

References

Abbas, M. J., Khan, M. A., Hamza, A., Alsenan, S., Rehman, A., Baili, J., & Zhang, Y. (2025). C3BAM-XAI: Convolutional Block Attention Module Enhanced Explainable Artificial Intelligence-Based Parkinson's Disease Stage Classification. *Cognitive Computation*, *17*(3), 1–11 https://doi.org/10.1007/s12559-025-10472-8

Altalib, A., McGregor, S., Li, C., & Perelli, A. (2025). Synthetic CT Image Generation From CBCT: A Systematic Review. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *9*(6), 691–707. https://doi.org/10.1109/trpms.2025.3533749

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. I., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263. https://doi.org/10.3322/caac.21834

Celard, P., Iglesias, E. I., Sorribes-Fdez, J. M., Romero, R. P. D., Vieira, A. S., & Borrajo, L. (2023). A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, *35*(3), 2291–2323. https://doi.org/10.1007/s00521-022-07953-4

Dembla, N., & Yadav, R. (2025). Early diagnosis of soybean disease using pretrained network: A comparative study of efficientnetv2 and mobilenetv2 with advanced feature extraction technique. *Grenze International Journal of Engineering & Technology*, *11*(1), 401–409.

Ebrahim Ghajari, N., & Fathi, A. (2024). Detection and classification of lung cancer in histopathology images using deep learning. *Journal of Computing and Security*, *11*(1), 19–28.

Faizi, M. K., Qiang, Y., Wei, Y., Qiao, Y., Zhao, J., Aftab, R., & Urrehman, Z. (2025). Deep learning-based lung cancer classification of CT images. *BMC Cancer*, *25*(1), 1–13. https://doi.org/10.1186/s12885-025-14320-8

Gupta, A., Kumar, A., & Rautela, K. (2024). Lung Cancer detection and classification using U-net and DARTS for medical CT images. *Multimedia Tools and Applications*, *84*(18), 19065–19085.

Hamdallak, M. (2025). IQ-OTHNCCD Lung Cancer Dataset. *The-IQ-OTHNCCD-Lung-Cancer-Dataset*. https://www. kaggle.com/datasets/hamdallak/

Hendriks, L. E. L., Remon, J., Faivre-Finn, C., Garassino, M. C., Heymach, J. V., Kerr, K. M., Tan, D. S.-W., Veronesi, G., & Reck, M. (2024). Non-small-cell lung cancer. *Nature Reviews Disease Primers*, *10*(1), 71. https://doi.org/10.1038/s41572-024-00551-9

Hussein, S., Cao, K., & Bagci, U. (2019). Risk stratification of lung nodules using 3d convolutional neural networks. *IEEE Access*, *7*, 134349–134360.

Jin, H., Yu, C., Zhang, J., Zheng, R., Fu, Y., & Zhao, Y. (2025). Multitask Swin Transformer for classification and characterization of pulmonary nodules in CT images. *Quantitative Imaging in Medicine and Surgery*, *15*(3), 1845–1861. https://doi.org/10.21037/qims-24-1619

Kanakarajan, K., Rajendran, S., & Acharya, U. R. (2026). Radiomics and deep learning fusion with clinical features for prognostic analysis of non-small cell lung cancer. *Journal of Digital Imaging*, *39*(2), 245–259.

Katar, O., Yildirim, O., Tan, R.-S., & Acharya, U. R. (2024). A Novel Hybrid Model for Automatic Non-Small Cell Lung Cancer Classification Using Histopathological Images. *Diagnostics*, *14*(22), 2497. https://doi.org/10.3390/diagnostics14222497

Kumar Lilhore, U., Simaiya, S., Sharma, Y. K., Kaswan, K. S., Brahma Rao, K. B. V., Maheswara Rao, V. V. R., Baliyan, A., Bijalwan, A., & Alroobaea, R. (2024a). A precise model for skin cancer diagnosis using hybrid U-Net and improved MobileNet-V3 with hyperparameters optimization. *Scientific Reports*, *14*(1), 1–23. https://doi.org/10.1038/s41598-024-54212-8

Kumar, S., Kumar, H., Kumar, G., Singh, S. P., Bijalwan, A., & Diwakar, M. (2024b). A methodical exploration of imaging modalities from dataset to detection through machine learning paradigms in prominent lung disease diagnosis: a review. *BMC Medical Imaging*, *24*(1), 30. https://doi.org/10.1186/s12880-024-01192-w

Li, Y., Yan, B., & He, S. (2023). Advances and challenges in the treatment of lung cancer. *Biomedicine & Pharmacotherapy*, *169*, 115891. https://doi.org/10.1016/j.biopha.2023.115891

Mahmoud, M., Wen, Y., Pan, X., Liufu, Y., & Guan, Y. (2025). Evaluation of recent lightweight deep learning architectures for lung cancer CT classification. *Frontiers in Oncology*, *15*, 1647701. https://doi.org/10.3389/fonc.2025.1647701

Majeed, A., Ruane, B., Shusted, C. S., Austin, M., Mirzozoda, K., Pimpinelli, M., Vojnika, J., Ward, L., Sundaram, B., Lakhani, P., Kane, G., Lev, Y., & Barta, J. A. (2022). Frequency of Statin Prescription Among Individuals with Coronary Artery Calcifications Detected Through Lung Cancer Screening. *American Journal of Medical Quality*, *37*(5), 388–395. https://doi.org/10.1097/jmq.0000000000000053

Navaneethakrishnan, M., Anand, M. V., Vasavi, G., & Rani, V. V. (2023). Deep Fuzzy SegNet-based lung nodule segmentation and optimized deep learning for lung cancer detection. *Pattern Analysis and Applications*, *26*(3), 1143–1159. https://doi.org/10.1007/s10044-023-01135-1

Peng, X., Dai, R., Ma, Y., Lin, B., Hui, X., Chen, X., & Lv, R. (2022). Early diagnosis and bioimaging of lung adenocarcinoma cells/organs based on spectroscopy machine learning. *Journal of Innovative Optical Health Sciences*, *15*(02), 1–12. https://doi.org/10.1142/s1793545822500110

Poch, E., Molina, A., & Piñeiro, G. (2022). Syndrome of inappropriate antidiuretic hormone secretion. *Medicina Clínica (English Edition)*, *159*(3), 139–146. https://doi.org/10.1016/j.medcle.2022.02.019

Rastogi, D., Johri, P., Donelli, M., Kumar, L., Bindewari, S., Raghav, A., & Khatri, S. K. (2025). Brain Tumor Detection and Prediction in MRI Images Utilizing a Fine-Tuned Transfer Learning Model Integrated Within Deep Learning Frameworks. *Life*, *15*(3), 327. https://doi.org/10.3390/life15030327

Raza, A., Ahmed, S., & Khan, M. A. (2026). Clinical validation of lightweight convolutional neural networks for ct-based lung cancer diagnosis. *Scientific Reports*, *16*(1), 1–14.

Saragih, G. S., Rustam, Z., & Aurelia, J. E. (2021). A hybrid model based on convolutional neural networks and fuzzy kernel K-medoids for lung cancer detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *24*(1), 126. https://doi.org/10.11591/ijeecs.v24.i1.pp126-133

Saxena, S., Prasad, S., Polnaya, A. M., & Agarwala, S. (2025). Hybrid deep convolution model for lung cancer detection with transfer learning. *Computer Science > Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.2501.02785

Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2022). Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier. *Neural Computing and Applications*, *34*(12), 9579–9592. https://doi.org/10.1007/s00521-020-04842-6

Sriramkumar, R., Selvakumar, K., & Jegan, J. (2025). Refining Chest X-ray Interpretation with Deep Transfer Learning Techniques. *Journal of Computer Science*, *21*(10), 2238–2255. https://doi.org/10.3844/jcssp.2025.2238.2255

Sun, Q., Zhang, Y., & Li, J. (2025). Radiomics and Deep Learning Fusion for Predicting Invasiveness of Lung Adenocarcinoma in Ground-Glass Nodules. *Scientific Reports*, *15*(1), 13447. https://doi.org/10.1038/s41598-025-13447-y

Vij, A. (2025). Multilevel Optimization with Hybrid Stack Model for Lung Cancer Classification. *Journal of Information Systems Engineering and Management*, *10*(28s), 993–1019. https://doi.org/10.52783/jisem.v10i28s.4762

Vij, A., & Kaswan, K. S. (2023). Prediction of Lung Cancer using Convolution Neural Networks. *Proceeding of the 2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, 737–741. https://doi.org/10.1109/aisc56616.2023.10085058

Zhou, L., Jain, A., Dubey, A. K., Singh, S. K., Gupta, N., Panwar, A., Kumar, S., Althaqafi, T. A., Arya, V., Alhalabi, W., & Gupta, B. B. (2025). FPA-based weighted average ensemble of deep learning models for classification of lung cancer using CT scan images. *Scientific Reports*, *15*(1), 19369. https://doi.org/10.1038/s41598-025-02015-w