International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Research Article)

Hybrid deep learning for interpretable lung cancer recognition across computed tomography and histopathological imaging modalities

Mohammad Rasel Mahmud [1], Hasib Fardin [2], Md Ismail Hossain Siddiqui [3], Anamul Haque Sakib [4] and Abdullah Al Sakib [5,*]

[1] Department of Management Information System, International American University, CA 90010, USA.
[2] Department of Engineering Management, Westcliff University, Irvine, CA 92614, USA.
[3] Department of Engineering/Industrial Management, Westcliff University, Irvine, CA 92614, USA.
[4] Department of Business Administration, International American University, Los Angeles, CA 90010, USA.
[5] Department of Information Technology, Westcliff University, Irvine, CA 92614, USA.

International Journal of Science and Research Archive, 2025, 15(01), 1798-1810

Publication history: Received on 13 March 2025; revised on 22 April 2025; accepted on 24 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1161

Abstract

Lung cancer is one of the deadliest cancers globally, primarily due to its silent development and difficulties with late diagnoses. Traditional diagnostic methods like manual CT scans and histopathological slide analysis face challenges such as observer variability, limited sensitivity, and difficulties in handling large volumes of data. Convolutional neural networks (CNNs) can automate image classification, but their limited receptive fields hinder complex tissue structure analysis. Vision Transformers (ViTs) provide a solution but typically need large datasets and significant computing power, making them impractical in clinical settings. Our study introduces a hybrid deep learning (DL) framework using the LEViT architecture to enhance lung cancer classification. We utilized two public datasets: IQ-OTH/NCCD, containing 1,097 CT images categorized into normal, benign, and malignant, and another dataset with 25,000 histopathological images across five tissue types. Our methodology included a multi-stage preprocessing pipeline to resize, reduce noise, enhance contrast, normalize, and augment data to tackle class imbalance and improve generalization. We evaluated our model using metrics like accuracy, F1 score, specificity, PR AUC, and Matthews Correlation Coefficient (MCC) through 10-fold stratified cross-validation. Our LEViT-based model surpassed top models such as CoAtNet and CrossViT, achieving 99.43% accuracy and 98.36% MCC on the IQ-OTH/NCCD dataset, and 99.02% accuracy with 97.97% MCC on the other dataset. Additionally, we developed a real-time web application for clinicians to upload images and receive visual explanations via Grad-CAM, promoting transparency in decision-making. This work provides a scalable, accurate, and explainable AI solution for lung cancer recognition, connecting high-performance algorithms with clinical practice.

Keywords: Lung cancer; Vision transformer; Medical imaging; Explainable AI; Diagnostic tool

1. Introduction

According to GLOBOCAN 2020, Lung cancer is the second most common cancer and the leading cause of cancer-related deaths globally, with approximately 2.2 million new cases and 1.8 million deaths each year [1]. The five-year survival rate is only 18.6%, dropping below 5% for late-stage diagnoses [2]. Non-small cell lung cancer (NSCLC) accounts for about 85% of cases and is often diagnosed at advanced stages due to minimal early symptoms [3]. Current diagnostic methods rely on visual inspections of CT scans and histopathological slides, which can be subjective and vary between observers, often lacking sensitivity for subtle tumor presentations. In resource-limited settings or high-throughput environments, this manual diagnostic approach is not scalable. Early and accurate detection is crucial for better patient outcomes and reducing healthcare system burdens. Traditional diagnostic methods, like radiographic interpretation,

[*] Corresponding author: Abdullah Al Sakib

often face issues such as inter-observer variability and limited sensitivity to subtle anomalies, especially in large-scale screenings [4]. This highlights the need for automated diagnostic systems that enhance identification accuracy and assist clinical decision-making. DL has advanced medical imaging notably, excelling in image recognition. CNNs are particularly effective for lung cancer identification in chest CT scans and histopathological slides.

While CNNs have achieved notable success, they are fundamentally constrained by their local receptive fields [5]. This limitation holds back their capacity to capture long-range dependencies and global contextual relationships, which are crucial for identifying irregular tumor patterns or subtle morphological differences in complex tissues. Furthermore, CNNs depend on large annotated datasets to generalize effectively [6], [7]. However, this is often a challenge in image classification applications due to issues related to data privacy and the high costs associated with data acquisition [8]. To overcome these issues, ViTs have emerged as an alternative to traditional CNNs, capturing global attention and contextual features in images [9]. However, ViTs require large amounts of data, are computationally intensive, and lack certain advantages of CNNs, like translation invariance and locality. Hybrid architecture that combines CNNs and ViTs has shown promise in addressing these issues. One notable model is LEViT, which effectively integrates CNN spatial feature extraction with ViT global attention in a computationally efficient way, making it ideal for resource-sensitive tasks like medical image analysis [10]. Furthermore, explainability in AI diagnostics is crucial for gaining acceptance in clinical settings [11]. While high-performance models can accurately predict diseases, their "black box" nature can reduce trust among clinicians. Techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) visualize key areas in input images that influence predictions, enhance transparency and build confidence in AI-assisted diagnoses.

In this study, we present a comprehensive framework for lung cancer classification that integrates the LEViT model with a robust preprocessing pipeline and an interpretable deployment environment (Figure 1). We focus on two diverse and widely used datasets: IQ-OTH/NCCD, which consists of CT scans categorized into normal, benign, and malignant classes, and another dataset featuring histopathology images of lung and colon tissue samples, divided into five classes. To address data imbalance and enhance model generalization, we implemented an extensive preprocessing and augmentation strategy that includes resizing, noise reduction, histogram equalization, and geometric transformations. The proposed model was trained using 10-fold stratified cross-validation and evaluated with five comprehensive metrics. Our key contributions are as follows:

- Proposed a LEViT-based hybrid DL model for multi-class lung cancer classification using both CT and histopathological images.
- Integrated Grad-CAM-based visual explanations to enhance model interpretability and clinical trust.
- Developed a real-time web application enabling clinicians to upload and analyze medical images with transparent AI outputs.
- Performed a comparative evaluation with state-of-the-art models, demonstrating the superiority of our approach across all key metrics.

The paper is structured as follows: Section 2 examines relevant literature to establish the basis for our study. Section 3 presents the datasets, methods, and architectural designs of the proposed models. In Section 4, we analyze the results, compare them with existing studies, and showcase our web application. Section 5 discusses the significance of our findings. Finally, section 6 concludes the article by proposing potential directions for future research.

2. Related Works

2.1. CNN Based Approaches

Raza et al. [12] developed Lung-EffNet, a transfer learning model using EfficientNetB1 for classifying multiple types of lung cancer based on 1,097 CT images from the IQ-OTH/NCCD dataset. The model achieved an accuracy of 99.10%. Data augmentation was used to tackle class imbalance, but the study faced challenges due to a small dataset size and limited external validation. Musthafa et al. [13] developed a double-layered CNN optimized with SMOTE and advanced preprocessing to classify lung cancer stages using 1,097 CT scans from the IQ-OTH/NCCD dataset. The model achieved 99.64% accuracy, but its generalizability is limited by demographic constraints and lack of external validation. Gopinath et al. [14] developed a Deep Fused Features-based Cat-Optimized Network (DFF-CON) for lung cancer classification using 78,090 CT images from the LIDC-IDRI dataset. The model, which combines saliency maps, CNNs, and Cat Swarm Optimization, achieved 99.92% accuracy. However, its practical use is limited due to reliance on synthetic augmentation and insufficient clinical validation.

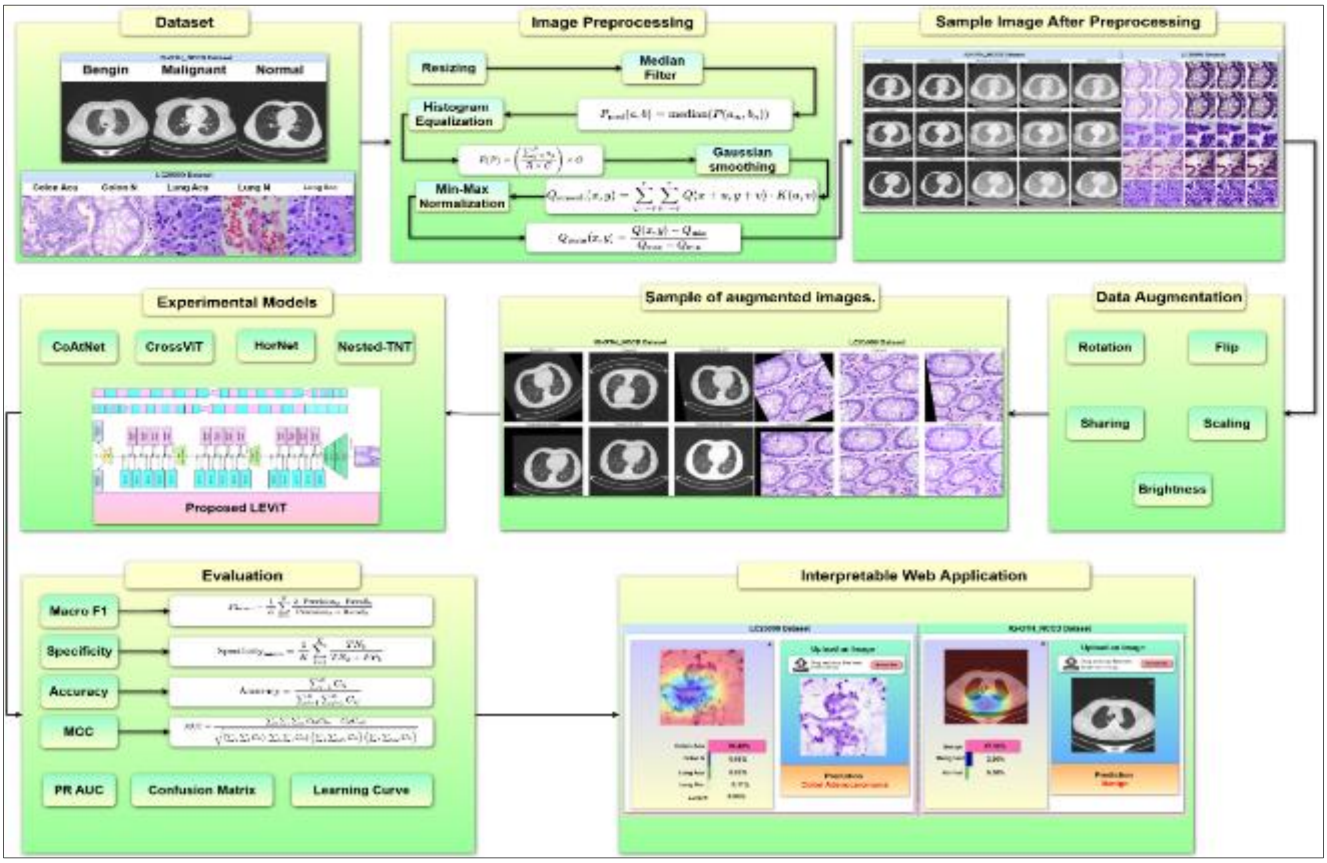

Figure 1 Overall methodology

Shah et al. [15] created a Deep Ensemble 2D CNN for detecting lung nodules, combining three different CNNs. Trained on 1,600 CT scans from the LUNA16 dataset, it achieved 95% accuracy. The development included thorough preprocessing and data augmentation. Nonetheless, it hasn't been tested on external datasets and doesn't account for 3D spatial information. Tyagi et al. [16] developed LCSCNet, a multi-level 3D DenseNet using concurrent squeeze-and-excitation and asymmetric convolutions to classify TNM-stage lung cancer. It was tested on the Lung-PET-CT Dx and NSCLC-Radiogenomics datasets, totaling 417 CT scans, achieving 97% accuracy. However, the model relies solely on CT input and has a limited dataset diversity, potentially hindering its generalization to broader clinical applications. Priya et al. [17] proposed SE-ResNeXt-50-CNN, integrating QDHE preprocessing and CNN classification for lung cancer detection using 3,552 CT images from the LUNA16 dataset. The model achieved 99.15% accuracy and outperformed existing models but requires significant computational resources and lacks external dataset validation.

2.2. ViT Based Approaches

Fu et al. [18] introduced LungMaxViT, a hybrid DL model combining CNNs with MaxViT-based attention to classify lung diseases from over 33,900 X-ray images (ChestX-ray14 and COVID-QU-Ex). Achieving 96.8% accuracy and 93.2% AUC, it excels in multi-class tasks. Nevertheless, class imbalance and computational intensity hinder broader scalability. Veasey et al. [19] proposed using LoRA-based fine-tuning on large vision models like SwinV2-b for lung nodule malignancy classification. They worked with the NLSTx and LIDC datasets, which included 857 and 647 nodules, respectively. Their best model improved the ROC AUC by 3% compared to previous methods and reduced parameters by 89.9%. However, it faced limitations, including modest clinical improvements and a lack of domain-specific data. Imran et al. [20] proposed a hybrid CNN-ViT hierarchical model for NSCLC classification using 15,000 histopathology images from the LC25000 dataset. The model achieved 98.8% accuracy, leveraging CNNs for local features and ViTs for global context. Moreover, the model faces challenges due to its reliance on a single dataset and the lack of validation in clinical deployment. Alsulami et al. [21] introduced LCCST-EMHI, ensemble model combining Swin Transformer features with BiLSTM-MHA, DDQN, and SSAE classifiers, optimized via the Walrus Optimization Algorithm. Tested on 25,000 histopathological images across five classes, it reached 98.92% accuracy. However, its real-world applicability remains unverified due to clinical deployment gaps.

Gulsoy et al. [22] developed FocalNeXt, a hybrid model combining ConvNeXt and FocalNet for lung cancer detection in CT scans. Tested on the IQ-OTH/NCCD dataset with 1,097 images, it achieved an accuracy of 99.81%. Despite its superior performance over existing ViT and CNN models, the model's real-world applicability is limited by dataset diversity and computational demands. Ko et al. [23] studied the impact of various optimizers on ViT models for classifying lung diseases with a dataset of 19,003 chest X-ray images. They found that the FastViT model using the NAdam optimizer reached an accuracy of 97.63%. While offering valuable insights into optimization strategies, the study's limitations include exclusion from CNN baselines and sensitivity to class imbalance.

3. Materials and Methods

3.1. Data Description

This study used two publicly available datasets: IQ-OTH/NCCD [24] and LC25000 [25]. The IQ-OTH/NCCD dataset consists of 1,097 CT scan images from 110 individuals, categorized into three groups: normal (55 subjects), benign (15), and malignant (40). These images were acquired using a Siemens SOMATOM CT scanner with a tube voltage of 120 kV, a slice thickness of 1 mm, and window settings ranging from 350 to 1200 Hounsfield Units (HU) for width and 50 to 600 HU for center. The distribution of labeled images includes 561 malignant, 120 benign, and 416 normal samples. The LC25000 dataset contains 25,000 high-resolution pathology images across five balanced categories: colon adenocarcinoma (Colon Aca), benign colonic tissue (Colon N), lung adenocarcinoma (Lung Aca), lung squamous cell carcinoma (Lung Scc), and benign lung tissue (Lung N), with 5,000 images per class. These images were digitized from pathology slides using high-resolution microscopes at 1024 x 768 pixels, ensuring detailed visual quality for DL applications. All images are de-identified and compliant with HIPAA regulations. Visual examples from both datasets are provided in Figure 2. For model development and evaluation, both datasets were divided into training, validation, and testing sets in an 80:5:15 ratio.

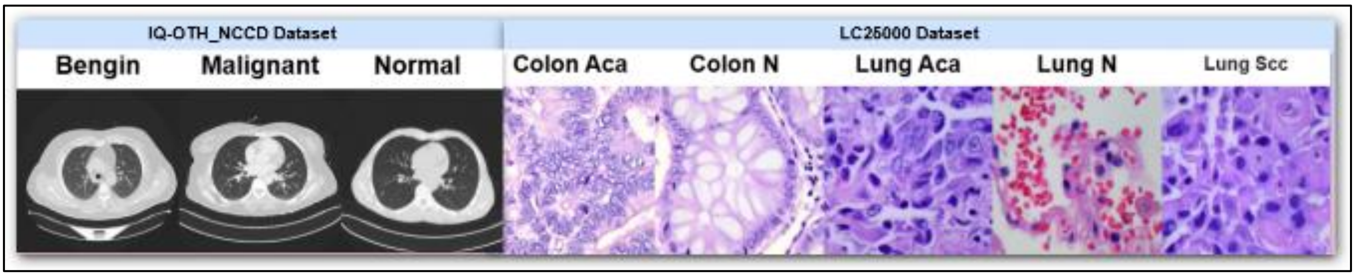

Figure 2 Sample image from each class from (a) IQ-OTH/NCCD and (b) LC25000 dataset

3.2. Data Preprocessing and Augmentation

To ensure uniformity across the dataset, all input images were resized to 224×224 pixels using Bilinear Interpolation, a technique that calculates each new pixel's value by taking a weighted average of the four closest neighboring pixels from the original image [26]. This ensures size consistency without significant loss of visual details. After resizing, a Median Filter was applied to remove salt-and-pepper noise while preserving edges [27]. This method replaces the pixel value at a given location with the median value from its surrounding neighborhood. If $\mathrm{P(a,b)}$ represents the pixel intensity at coordinates $\mathrm{(a,b)}$, the filtered intensity $\mathrm{P_{med}(a,b)}$ is computed as Equation 1, where $\mathrm{(a_m,b_n)}$ are the pixel coordinates are within the kernel window centered at $\mathrm{(a,b)}$. This helps reduce noise while preserving important structural features.

$$P_{\text{med}}(a,b) = \text{median}\big(P(a_m,b_n)\big) \tag{1}$$

We used Histogram Equalization to improve contrast and highlight texture differences in the image. This method redistributes intensity values across the dynamic range, enhancing the image's separability. Let ($\mathrm{h(P)}$) represent the histogram of pixel intensities, and ($\mathrm{s_q}$) be the cumulative frequency up to intensity level (P). For an image with $R \times C$ pixels and a maximum intensity level of (G), the transformation function $E(P)$ is defined as Equation 2.

$$E(P) = \left(\frac{\sum_{q=0}^{P} s_q}{R \times C}\right) \times G \tag{2}$$

This operation enhances local contrast and makes low contrast features more prominent.

Gaussian smoothing was used to reduce high-frequency noise while preserving important image structures. The smoothed intensity $Q_{\text{smooth}}(x, y)$ at coordinates (x,y) is calculated by convolving the image with a Gaussian kernel K(u, v), as defined by Equation 3, where (r) is the kernel's radius. This method effectively suppresses noise while maintaining the image's overall structure.

$$Q_{\text{smooth}}(x, y) = \sum_{u=-r}^{r} \sum_{v=-r}^{r} Q(x + u, y + v) \cdot K(u, v) \tag{3}$$

Min-Max Normalization was utilized to standardize the input values for training. Let $Q_{\text{min}} and Q_{\text{max}}$ represent the minimum and maximum pixel intensity values in the image. The normalized pixel intensity $Q_{\text{norm}}(x, y)$ is computed as shown in Equation 4. This normalization ensures consistent value ranges, improving model convergence. Figure 3 demonstrate preprocessed examples from the LC25000 and IQOTH/NCCD datasets.

$$Q_{\text{norm}}(x, y) = \frac{Q(x, y) - Q_{min}}{Q_{max} - Q_{min}} \tag{4}$$

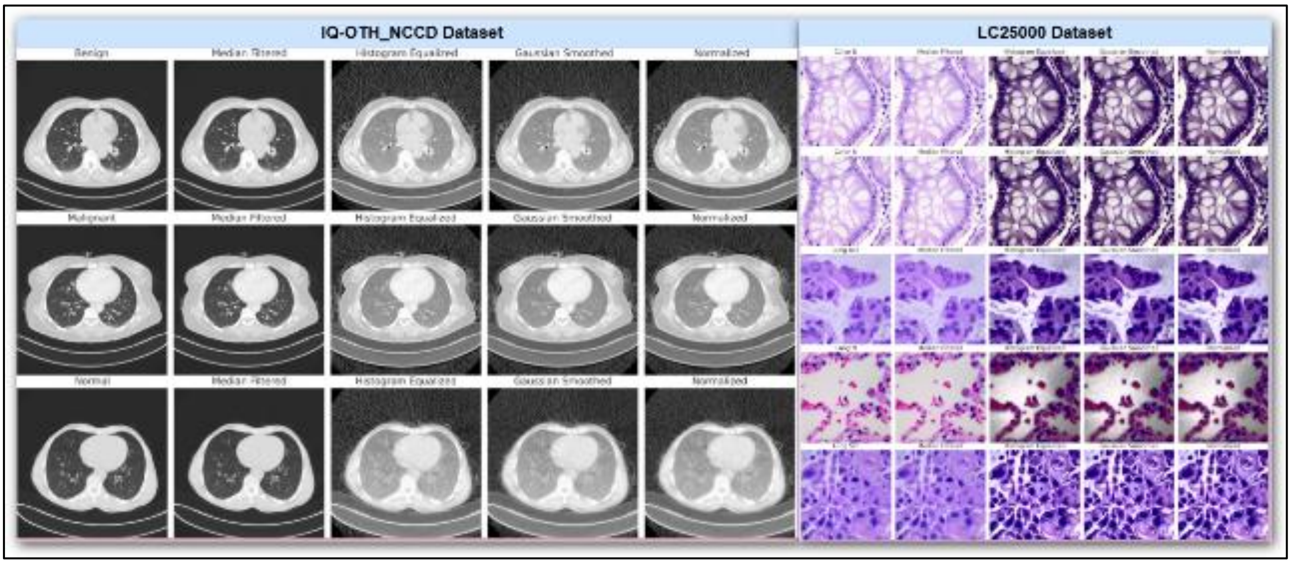

Figure 3 Sample images from each class of both IQ-OTH/NCCD and LC25000 dataset after preprocessing

To improve the model's generalization and adaptability, we used data augmentation techniques only training set to increase variability and achieve class balance. For the IQ-OTH/NCCD dataset, we applied oversampling, resulting in 561 images per class (normal, benign, and malignant), raising the total from 1,097 to 1,683 images. Image augmentation included random rotations of ±25°, horizontal and vertical flips, ±0.15 distortion, and translations up to 12% of the image size. We also applied scaling transformations between 0.85 and 1.15 and brightness adjustments ranging from 0.75 to 1.25. The LC25000 dataset, already balanced with 5,000 images per class (totaling 25,000 images), received the same augmentation methods. This provided 252 test images for IQ-OTH/NCCD and 3,750 for LC25000. Transformations are illustrated in Figure 4.

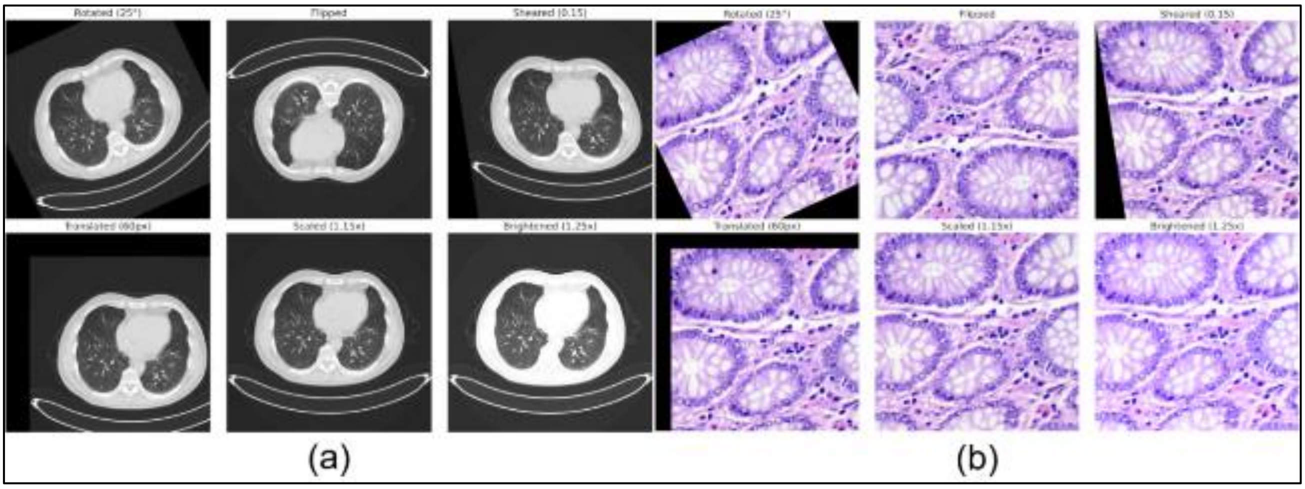

Figure 4 Sample images from both (a) IQ-OTH/NCCD and (b) LC25000 dataset after augmentation

3.3. Experimental Models

3.3.1. ViT Models

Transfer learning has become a vital technique in medical image analysis, particularly for tasks like lung cancer identification, where large, annotated datasets are often scarce. In this study, we employed state-of-the-art DL models—CoAtNet, Nested-TNT, CrossViT, and HorNet—using transfer learning for lung cancer classification. These models were selected for their architectural strengths in capturing both local texture and global structural features from medical images. Each model introduces innovative mechanisms that address the limitations of traditional CNNs and enhance the capability to learn from high-resolution clinical imagery.

CoAtNet combines convolutional operations with self-attention mechanisms, unifying the generalization capability of CNNs with the scalability of Transformers. This combination is particularly effective for local feature extraction and contextual understanding [28]. Nested-TNT utilizes a Transformer-in-Transformer design, allowing it to capture fine-grained details within each patch while also modeling dependencies between patches [29]. This hierarchical representation is beneficial for identifying small lesions and heterogeneous cancerous tissues. CrossViT enhances visual representation by processing input images at multiple scales using a dual-branch ViT. This multi-scale attention mechanism allows the model to better detect abnormalities that vary in size, shape, and location [30]—common challenges in lung cancer diagnosis. On the other hand, HorNet introduces high-order spatial interactions through recursive convolutional attention, delivering strong performance while maintaining computational efficiency [31]. This makes it particularly suitable for real-time or resource-constrained clinical applications.

3.3.2. Proposed LEViT

The LEViT architecture is a hybrid visual model that combines CNNs with ViTs. This design (Figure 5) provides a robust, lightweight, and high-speed framework for lung cancer identification. Its hybrid nature is particularly effective in extracting both low-level spatial and high-level semantic features from medical images. The input to the LEViT model consists of RGB lung image samples resized to 224 × 224 pixels. The model starts with a convolutional stem made up of four consecutive 3 × 3 convolution layers. These layers reduce the spatial dimensions while increasing the number of feature channels, resulting in a representation of 256 × 14 × 14. This feature map is then tokenized and processed through stacked attention-based transformer stages. Each stage in the transformer block includes Multi-Head Self-Attention (MHSA) and two-layer Multi-Layer Perceptrons (MLPs). The MHSA allows the model to capture complex relationships between image patches by projecting the input sequence into multiple attention heads [32].

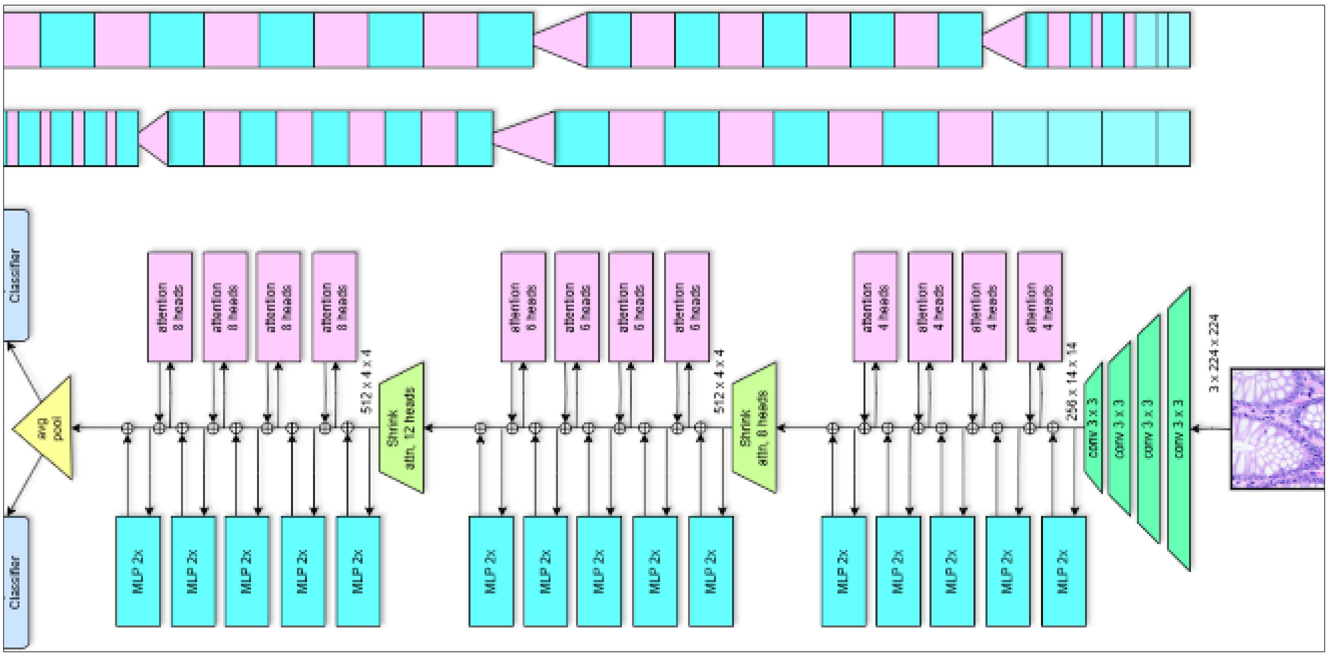

Figure 5 Overview of the proposed LEViT architecture that combines convolutional layers for early spatial feature extraction with hierarchical transformer blocks for capturing global dependencies

The MHSA operation is defined by Equations 5-6, where Q, K, and V represent the query, key, and value matrices derived from the input embeddings. $W_i^Q,\ W_i^K, and\ W_i^V$ are learnable projection matrices for each head.

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \tag{5}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6}$$

After each attention block, there is a two-layer multi-layer perceptron (MLP) that utilizes a non-linear activation function. This design adds depth and allows the model to learn complex transformations [33]. The operation of the MLP is described by Equation 7, where $W_1\ and\ W_2$ are learnable weights, $b_1\ and\ b_2$ are biases, and σ represents a non-linear activation function, typically GELU or ReLU.

$$\text{MLP}(x) = W_2 \cdot \sigma(W_1 \cdot x + b_1) + b_2 \tag{7}$$

LEViT employs a hierarchical structure featuring attention blocks that gradually downsample the spatial resolution while increasing the channel dimension [34]. The model starts with attention layers that have 4 heads and progresses to 6, 8, and ultimately 12 heads as it becomes deeper. This design enhances the network's ability to capture increasingly abstract and global features. The final output is a compact tensor with dimensions $512\ \times 4\ \times 4$. An average pooling layer is then applied to aggregate the spatial features into a 512-dimensional feature vector, which is subsequently passed to a supervised classifier for the final prediction. This architecture ensures both discriminative power and computational efficiency.

3.4. Training Parameters and Evaluation

The training process consisted of 30 epochs, with each epoch representing a complete pass through the dataset. The Adam optimizer was utilized to adjust the model parameters, with a fixed learning rate of 0.001. Categorical cross-entropy served as the loss function. To prevent overfitting and enhance generalization, we implemented callback functions such as early stopping, a learning rate scheduler, and model checkpointing. Model performance was measured using accuracy, F1 score, specificity, PR AUC, and MCC. Accuracy assessed overall correct predictions, while the F1 score emphasized the balance between precision and recall, which is important for imbalanced datasets. Specificity focused on correctly identifying negative cases to reduce false positives, PR AUC analyzed the precision-recall trade-off, and MCC provided a comprehensive performance measure based on all confusion matrix components. To ensure a robust evaluation, 10-fold stratified cross-validation was employed. The dataset was divided into ten subsets while maintaining class distribution. In each iteration, one subset was used for validation, and the remaining nine were used for training, repeating this process for each subset.

4. Results and Discussion

The results show the models' performance across 3-class (IQ-OTH/NCCD) and 5-class (LC25000) datasets before augmentation. LeViT consistently outperforms all the other models in both datasets (Table 1). For the IQ-OTH/NCCD dataset, it achieves the highest accuracy at 97.88% and an MCC of 96.69%. In the LC25000 dataset, it again leads with an F1 score of 97.13% confirming its robustness and reliability.

Table 1 Performance comparison of models in both LC25000 and IQ-OTH/NCCD dataset before augmentation

Model	Dataset	Accuracy	F1	Specificity	PR AUC	MCC
LEViT	3-class	97.88 ± 0.34	96.93 ± 0.25	98.03 ± 0.37	98.17 ± 0.29	96.69 ± 0.47
HorNet		96.66 ± 0.50	95.03 ± 0.45	96.36 ± 0.54	96.16 ± 0.35	95.28 ± 0.30
CoAtNet		94.80 ± 0.43	95.30 ± 0.79	95.14 ± 0.39	95.39 ± 0.64	93.79 ± 0.28
Nested-TNT		93.44 ± 1.04	91.60 ± 0.63	96.39 ± 0.90	94.69 ± 0.46	93.33 ± 0.18
CrossViT		92.26 ± 1.51	90.69 ± 0.98	92.81 ± 0.58	92.87 ± 0.90	91.89 ± 0.41
LEViT	5-class	98.29 ± 0.31	97.13 ± 0.25	97.96 ± 0.19	98.34 ± 0.34	96.75 ± 0.05
CoAtNet		97.14 ± 0.40	96.73 ± 0.49	96.97 ± 0.21	97.27 ± 0.37	96.15 ± 0.24
HorNet		96.40 ± 0.46	96.12 ± 0.16	96.37 ± 0.41	95.84 ± 0.47	94.76 ± 0.34
CrossViT		94.88 ± 0.24	94.33 ± 0.56	95.86 ± 0.69	95.38 ± 0.45	93.57 ± 1.03
Nested-TNT		94.47 ± 1.07	94.87 ± 0.69	93.03 ± 0.84	95.24 ± 0.45	94.20 ± 0.57

CoAtNet closely follows in the IQ-OTH/NCCD dataset, with solid specificity at 95.14%. In the LC25000 dataset, it achieves a PR AUC of 97.27%, making it a dependable model just behind LeViT. HorNet performs well but slightly trails CoAtNet, it shows good F1 at 95.03% in the IQ-OTH/NCCD. In the LC25000 setting, it scores 94.76% MCC, indicating solid performance, but not at the top level. CrossViT underperforms in the IQ-OTH/NCCD dataset, recording the lowest MCC at 91.89%. Its performance improves in the 5-class dataset with 93.57% MCC, but it still ranks behind the top models. Nested-TNT outperforms CrossViT in both datasets, achieving 93.44% accuracy and 94.47% accuracy though it still lags behind other models.

Data augmentation led to substantial performance improvements for all models in Table 2 on both the 3-class IQ-OTH/NCCD dataset and the 5-class LC25000 dataset. LeViT achieved the highest accuracy, 99.43% for IQ-OTH/NCCD and 99.02% for LC25000. HorNet and CoAtNet followed, while CrossViT (93.70%) and Nested-TNT had the lowest accuracies. LeViT also led in F1 Score with 98.44% for IQ-OTH/NCCD, and HorNet scored 97.76% for LC25000. CoAtNet had competitive F1 scores, but Nested-TNT performed the weakest in both datasets. In specificity, LeViT scored 99.57% on IQ-OTH/NCCD and 99.12% on LC25000. Nested-TNT did well on IQ-OTH/NCCD but struggled with LC25000. CoAtNet and CrossViT showed consistent performance but did not lead. LeViT had the highest PR AUC with 99.94% for IQ-OTH/NCCD and 99.18% for LC25000, while CrossViT had the lowest at 94.87% in the 3-class dataset. LeViT's performance was confirmed by MCC scores of 98.36% and 97.97% for the two datasets, with CoAtNet and HorNet following, and CrossViT scoring the lowest at 93.71% in the 3-class dataset.

Table 2 Performance comparison of models in both LC25000 and IQ-OTH/NCCD dataset after augmentation

Model	Dataset	Accuracy	F1	Specificity	PR AUC	MCC
LEViT	3-class	99.43 ± 0.12	98.44 ± 0.69	99.57 ± 0.29	99.94 ± 0.51	98.36 ± 0.09
HorNet		98.31 ± 0.97	96.70 ± 0.05	98.30 ± 0.55	97.17 ± 0.30	96.56 ± 0.14
CoAtNet		96.30 ± 0.84	96.41 ± 0.95	96.31 ± 0.51	97.35 ± 0.14	95.13 ± 0.03
Nested-TNT		94.79 ± 0.39	93.49 ± 0.25	97.41 ± 0.19	95.73 ± 0.42	94.78 ± 0.31
CrossViT		93.70 ± 1.33	92.48 ± 1.11	94.45 ± 1.44	94.87 ± 1.20	93.71 ± 1.61
LEViT		99.02 ± 0.98	98.32 ± 0.90	99.12 ± 0.22	99.18 ± 0.31	97.97 ± 0.57

CoAtNet	5-class	98.73 ± 0.47	98.47 ± 0.66	98.74 ± 0.28	99.29 ± 0.87	97.94 ± 1.07
HorNet		98.26 ± 0.12	97.76 ± 0.44	97.54 ± 0.34	97.25 ± 0.25	96.67 ± 0.31
CrossViT		96.15 ± 1.09	96.06 ± 1.70	97.11 ± 1.29	95.33 ± 1.17	94.64 ± 1.47
Nested-TNT		96.09 ± 1.27	96.22 ± 1.49	94.45 ± 1.18	96.97 ± 1.14	95.47 ± 1.03

Figure 6 illustrates the training and validation learning curves over 30 epochs for both datasets. For the IQ-OTH_NCCD dataset, training and validation loss decrease consistently, approaching zero by epoch 25, indicating effective learning without overfitting. Minor fluctuations in validation loss remain stable. Both accuracy curves rise steadily, nearly reaching 99% by the final epoch, demonstrating strong generalization and consistent performance. Recall and precision also show upward trends, stabilizing above 97%, indicating the model effectively detects true positives and minimizes false positives. For the LC25000 dataset, similar trends are observed. The loss curves decline sharply, converging to nearly zero, while accuracy quickly climbs and plateaus around 99%, confirming robust generalization. Recall and precision improve steadily, achieving nearly perfect values by the end of training, with high confidence and minimal errors.

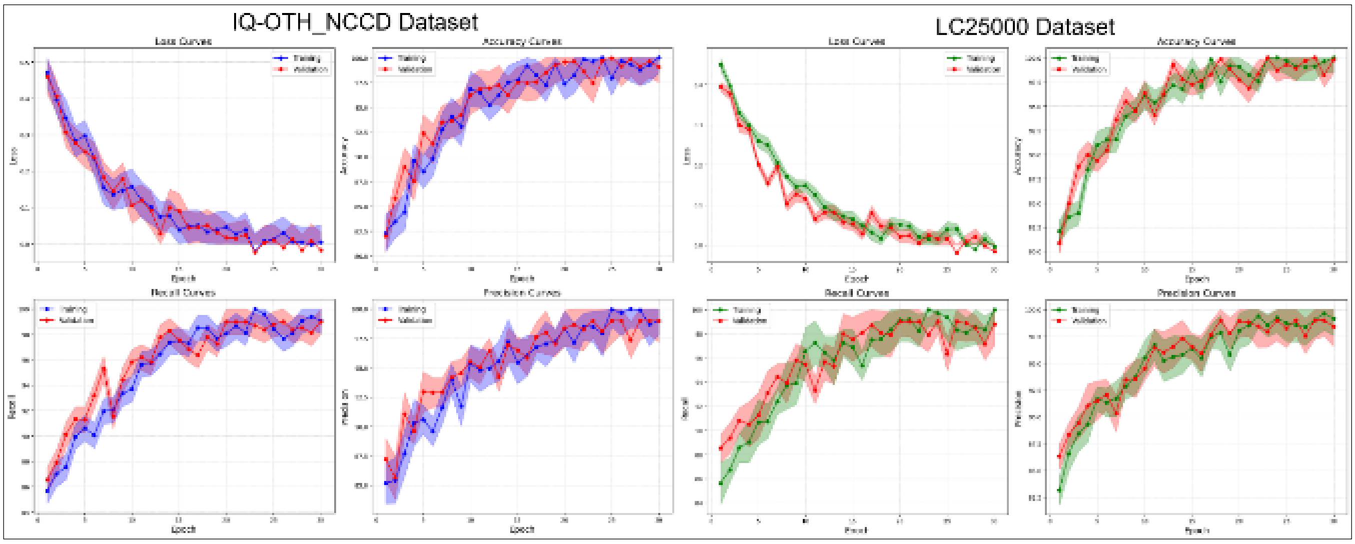

Figure 6 Learning curve of the proposed LEViT model for both datasets

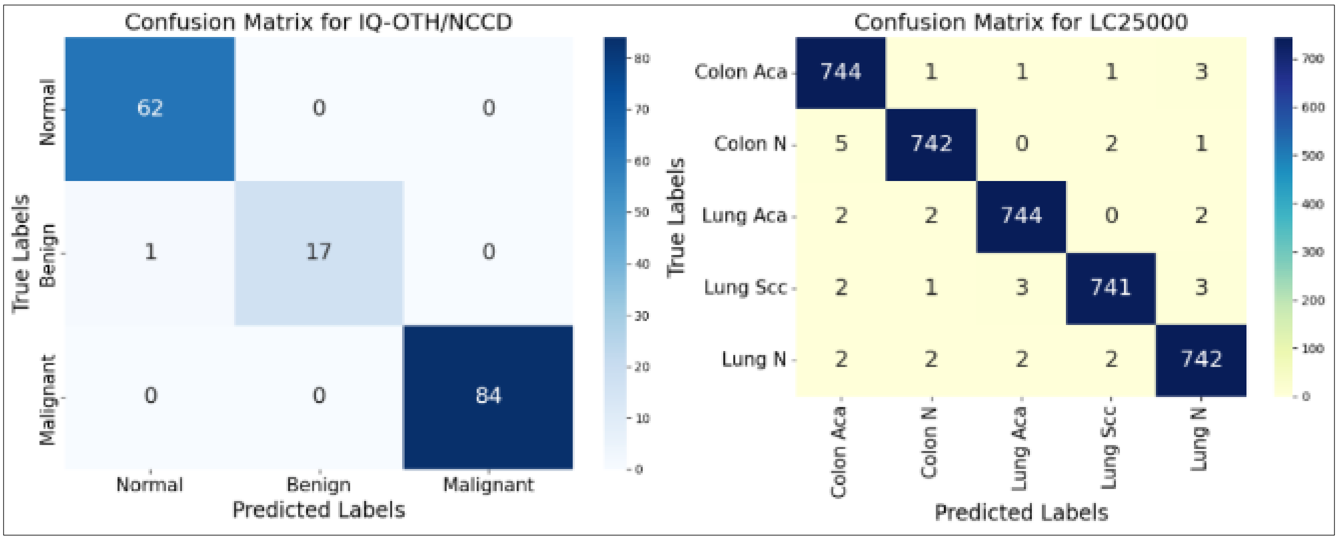

Figure 7 Confusion matrices of the proposed LEViT model for both datasets

Figure 7 shows the confusion matrices for the LEViT model applied to the IQ-OTH/NCCD and LC25000 datasets, demonstrating its strong classification performance. On the IQ-OTH/NCCD dataset, the model achieved perfect accuracy, correctly classifying all Normal (62/62) and Malignant (84/84) samples, with only one Benign sample misclassified as Normal. For the LC25000 dataset, which includes five classes—Colon Aca, Colon Normal, Lung Adenocarcinoma, Lung Squamous Cell Carcinoma, and Lung Normal—the model correctly classified at least 741 out of 750 samples in each category. Minor misclassifications occurred primarily between similar categories, such as Colon Aca and Colon Normal, or among the Lung subtypes, which is expected due to their visual similarities.

Figure 8 shows a web-based application for medical image classification with explainable AI (XAI). In the left panel, a histopathological image from the LC25000 dataset is uploaded, resulting in a 98.42% confidence level for Colon Aca. Other classes receive probabilities below 1%. The Grad-CAM visualization indicates key areas, highlighted in blue and yellow, that influenced the classification, focusing on dense cellular regions. In the right panel, a chest CT scan from the IQ-OTH/NCCD dataset is processed, predicting a Benign case with a 97.32% confidence score. Minor probabilities are assigned to Malignant (2.30%) and Normal (0.38%). The Grad-CAM output emphasizes the thoracic region, particularly around the lungs.

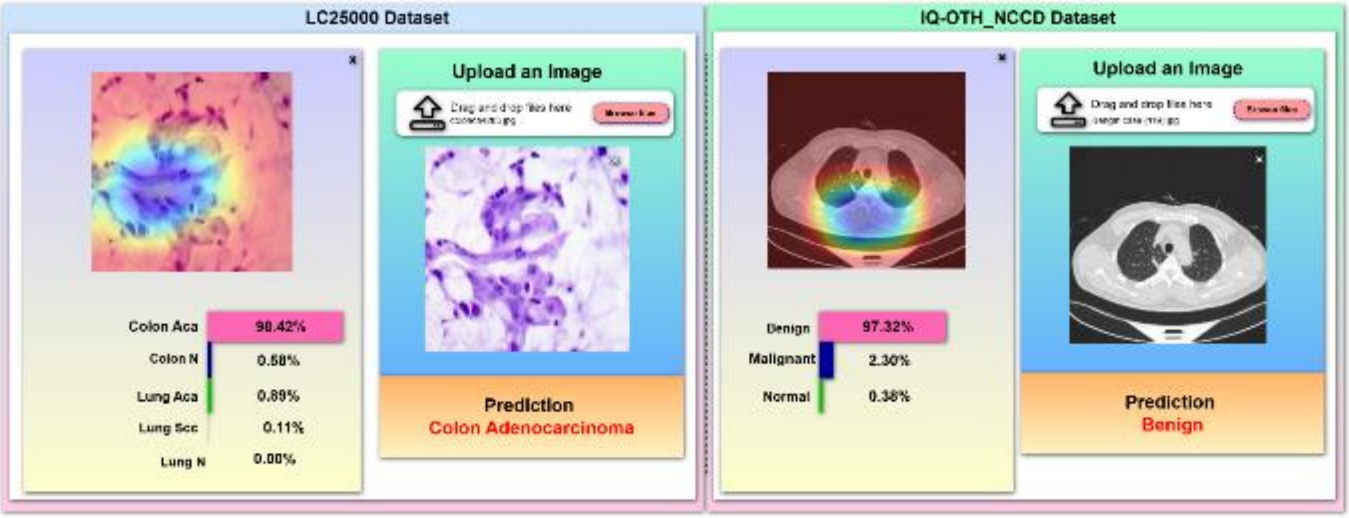

Figure 8 Web application showcasing explainable AI predictions using Grad-CAM

Table 3 compares existing methods with our proposed LeViT approach. The model achieves impressive accuracy, scoring 99.43% on the IQ-OTH/NCCD dataset (3-class) and 99.02% on the LC25000 dataset (5-class), outperforming all other models. For context, earlier methods like SE-ResNeXt-50 and EfficientNetB1 achieved 99.15% and 99.10% accuracy, respectively, while CNN-ViT Hybrid and LCCST-EMHI reached up to 98.92%. Additionally, our model is practical for real-world use as it is deployed as a real-time web application, allowing clinicians to upload images and receive instant predictions. We also incorporate explainable AI (XAI), using Grad-CAM for visualizing decision-making areas in medical images, enhancing transparency and trustworthiness, which is crucial in clinical settings.

Our proposed LEViT model outperformed existing architectures thanks to its hybrid design, which integrates convolutional stems for local feature extraction with hierarchical transformer stages for long-range dependencies. It employs progressive multi-head self-attention (increasing from 4 to 12 heads) for multi-scale representation learning, essential for detecting dispersed lesions in CT scans and diverse patterns in histopathology. LEViT strikes a balance between representational power and computational efficiency, enabling quicker inference without compromising accuracy. The preprocessing pipeline includes histogram equalization, Gaussian smoothing, and median filtering to enhance contrast, reduce noise, and maintain structural details. Data augmentation techniques improved model accuracy by 1.55% and increased the MCC by 1.67%, demonstrating better generalization and robustness. The web application provides real-time predictions and uses Grad-CAM visualizations to highlight important areas in images. It detects lung masses in CT scans and identifies abnormal nuclei in histopathology, enhancing interpretability and building trust among clinicians. LEViT's lightweight architecture allows easy deployment on mid-range GPUs for real-time integration into PACS and HIS. Its compatibility with both CT and histology supports cross-departmental use, and the explainable interface complies with clinical and regulatory standards for AI-assisted diagnostics.

Table 3 Comparative analysis of existing models and the proposed LeViT-based approach

Model	Dataset	Data	Accuracy	Application	XAI
Lung-EffNet [12]	IQ-OTH/NCCD	1,097	99.10%	No	No
Deep Ensemble [15]	LUNA16	1,600	95%	No	No
LCSCNet [16]	Lung-PET-CT Dx + NSCLC	417	97%	No	No
SE-ResNeXt-50-CNN [17]	LUNA16	3,552	99.15%	No	No
LungMaxViT [18]	ChestX-ray14 + COVID-QU-Ex	33,900+	96.80%	No	Yes
CNN-ViT Hybrid [20]	LC25000	15,000	98.80%	No	No
LCCST-EMHI [21]	Histopathology Images	25,000	98.92%	No	No
FocalNeXt [22]	IQ-OTH/NCCD	1,097	99.81%	No	No
CrossViT [23]	Chest X-ray multi-source	19,003	97.63%	No	No
Proposed LEViT (Ours)	IQ-OTH/NCCD	1,683	99.43%	Yes	Yes
	LC25000	25,000	99.02%	Yes	Yes

Despite their high performance, resizing fixed-size inputs can result in a loss of spatial detail in high-resolution pathology images, potentially making deep attention layers less sensitive to positional information and affecting the recognition of modest cues. Domain shifts between institutions pose challenges due to different imaging techniques and demographics. Besides, Grad-CAM's inability to quantify uncertainty limits its interpretability in unclear cases. Future enhancements should focus on multi-resolution transformers, deformable attention mechanisms, and integrating vision-language models. Improving uncertainty estimation through methods like Bayesian inference and advanced explainable AI techniques, including Layer-wise Relevance Propagation (LRP) and counterfactuals, will increase reliability. Validation on cross-institutional datasets and exploration of federated learning will support privacy-preserving and scalable solutions.

5. Conclusion

Detecting and treating lung cancer is challenging due to late-stage diagnoses and complex classification. This research presents a hybrid LEViT-based vision model integrated into a real-time web application for early lung cancer recognition. The framework tackles issues like class imbalance, model generalization, interpretability, and clinical deployment. By combining convolutional operations with hierarchical self-attention, the model balances scalability, accuracy, and efficiency. The web tool enhances access to diagnostic support, making it suitable for various healthcare settings, from urban centers to rural clinics and telemedicine platforms. While the system performs well in controlled experiments, it still needs extensive validation in real-world clinical settings. Factors like patient diversity and hardware limitations could impact results. Its high computational demands may also limit use in resource-poor facilities. Future efforts should focus on validating performance with multi-center datasets, incorporating various imaging techniques, and improving compatibility across platforms for broader healthcare use. Developing interpretable and accessible AI systems can enhance diagnostic accuracy and support earlier interventions, leading to improved patient outcomes.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

References

[1] A. Leiter, R. R. Veluswamy, and J. P. Wisnivesky, “The global burden of lung cancer: current status and future trends,” Nature Reviews Clinical Oncology 2023 20:9, vol. 20, no. 9, pp. 624–639, Jul. 2023, doi: 10.1038/s41571-023-00798-3.

[2] G. Li, X. Zhu, J. Liu, S. Li, and X. Liu, “Metal Oxide Semiconductor Gas Sensors for Lung Cancer Diagnosis,” Chemosensors 2023, Vol. 11, Page 251, vol. 11, no. 4, p. 251, Apr. 2023, doi: 10.3390/CHEMOSENSORS11040251.

[3] H. Padinharayil et al., “Non-small cell lung carcinoma (NSCLC): Implications on molecular pathology and advances in early diagnostics and therapeutics,” Genes Dis, vol. 10, no. 3, pp. 960–989, May 2023, doi: 10.1016/J.GENDIS.2022.07.023.

[4] N. Anan, R. Zainon, and M. Tamal, “A review on advances in 18F-FDG PET/CT radiomics standardisation and application in lung disease management,” Insights Imaging, vol. 13, no. 1, pp. 1–22, Dec. 2022, doi: 10.1186/S13244-021-01153-9/FIGURES/11.

[5] Y. Ma, M. Yu, H. Lin, C. Liu, M. Hu, and Q. Song, “Efficient analysis of deep neural networks for vision via biologically-inspired receptive field angles: An in-depth survey,” Information Fusion, vol. 112, p. 102582, Dec. 2024, doi: 10.1016/J.INFFUS.2024.102582.

[6] R. Haque et al., “Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning,” BioMedInformatics 2024, Vol. 4, Pages 966-991, vol. 4, no. 2, pp. 966–991, Apr. 2024, doi: 10.3390/BIOMEDINFORMATICS4020054.

[7] A. Al-Sakib, F. Islam, R. Haque, M. B. Islam, A. Siddiqua, and M. M. Rahman, “Classroom Activity Classification with Deep Learning,” 2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024, 2024, doi: 10.1109/ICICACS60521.2024.10498187.

[8] M. D. Hosen et al., “Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning,” 2024 6th International Conference on Electrical Engineering and Information &amp; Communication Technology (ICEEICT), pp. 1163–1168, May 2024, doi: 10.1109/ICEEICT62016.2024.10534322.

[9] A. Khan et al., “A survey of the vision transformers and their CNN-transformer based variants,” Artificial Intelligence Review 2023 56:3, vol. 56, no. 3, pp. 2917–2970, Oct. 2023, doi: 10.1007/S10462-023-10595-0.

[10] B. Li, J. Wang, Z. Guo, and Y. Li, “Automatic detection of schizophrenia based on spatial–temporal feature mapping and LeViT with EEG signals,” Expert Syst Appl, vol. 224, p. 119969, Aug. 2023, doi: 10.1016/J.ESWA.2023.119969.

[11] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective,” BMC Med Inform Decis Mak, vol. 20, no. 1, pp. 1–9, Dec. 2020, doi: 10.1186/S12911-020-01332-6/PEER-REVIEW.

[12] R. Raza et al., “Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images,” Eng Appl Artif Intell, vol. 126, p. 106902, Nov. 2023, doi: 10.1016/J.ENGAPPAI.2023.106902.

[13] M. M. Musthafa, I. Manimozhi, T. R. Mahesh, and S. Guluwadi, “Optimizing double-layered convolutional neural networks for efficient lung cancer classification through hyperparameter optimization and advanced image pre-processing techniques,” BMC Med Inform Decis Mak, vol. 24, no. 1, pp. 1–21, Dec. 2024, doi: 10.1186/S12911-024-02553-9/FIGURES/13.

[14] A. Gopinath, P. Gowthaman, M. Venkatachalam, and M. Saroja, “Computer aided model for lung cancer classification using cat optimized convolutional neural networks,” Measurement: Sensors, vol. 30, p. 100932, Dec. 2023, doi: 10.1016/J.MEASEN.2023.100932.

[15] A. A. Shah, H. A. M. Malik, A. H. Muhammad, A. Alourani, and Z. A. Butt, “Deep learning ensemble 2D CNN approach towards the detection of lung cancer,” Scientific Reports 2023 13:1, vol. 13, no. 1, pp. 1–15, Feb. 2023, doi: 10.1038/s41598-023-29656-z.

[16] S. Tyagi and S. N. Talbar, “LCSCNet: A multi-level approach for lung cancer stage classification using 3D dense convolutional neural networks with concurrent squeeze-and-excitation module,” Biomed Signal Process Control, vol. 80, p. 104391, Feb. 2023, doi: 10.1016/J.BSPC.2022.104391.

[17] A. Priya and P. Shyamala Bharathi, “SE-ResNeXt-50-CNN: A deep learning model for lung cancer classification,” Appl Soft Comput, vol. 171, p. 112696, Mar. 2025, doi: 10.1016/J.ASOC.2025.112696.

[18] X. Fu, R. Lin, W. Du, A. Tavares, and Y. Liang, “Explainable hybrid transformer for multi-classification of lung disease using chest X-rays,” Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 1–19, Feb. 2025, doi: 10.1038/s41598-025-90607-x.

[19] B. P. Veasey and A. A. Amini, “Low-Rank Adaptation of Pre-trained Large Vision Models for Improved Lung Nodule Malignancy Classification,” IEEE Open J Eng Med Biol, 2025, doi: 10.1109/OJEMB.2025.3530841.

[20] M. Imran, B. Haq, E. Elbasi, A. E. Topcu, and W. Shao, “Transformer Based Hierarchical Model for Non-Small Cell Lung Cancer Detection and Classification,” IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3449230.

[21] A. A. Alsulami, A. Albarakati, A. A. M. AL-Ghamdi, and M. Ragab, “Identification of Anomalies in Lung and Colon Cancer Using Computer Vision-Based Swin Transformer with Ensemble Model on Histopathological Images,” Bioengineering 2024, Vol. 11, Page 978, vol. 11, no. 10, p. 978, Sep. 2024, doi: 10.3390/BIOENGINEERING11100978.

[22] T. Gulsoy and E. Baykal Kablan, “FocalNeXt: A ConvNeXt augmented FocalNet architecture for lung cancer classification from CT-scan images,” Expert Syst Appl, vol. 261, p. 125553, Feb. 2025, doi: 10.1016/J.ESWA.2024.125553.

[23] J. Ko, S. Park, and H. G. Woo, “Optimization of vision transformer-based detection of lung diseases from chest X-ray images,” BMC Med Inform Decis Mak, vol. 24, no. 1, pp. 1–8, Dec. 2024, doi: 10.1186/S12911-024-02591-3/FIGURES/3.

[24] hamdalla alyasriy, “The IQ-OTHNCCD lung cancer dataset,” vol. 1, 2020, doi: 10.17632/BHMDR45BH2.1.

[25] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, “Lung and Colon Cancer Histopathological Image Dataset (LC25000),” Dec. 2019, Accessed: Mar. 30, 2025. [Online]. Available: https://arxiv.org/abs/1912.12142v1

[26] M. S. Rahman et al., “Effective Disease Recognition in Cucumbers: A Web-Based Application Using Transfer Learning Models,” 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings, pp. 59–64, 2024, doi: 10.1109/RAAICON64172.2024.10928353.

[27] R. Haque et al., “A Scalable Solution for Pneumonia Diagnosis: Transfer Learning for Chest X-ray Analysis,” 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), pp. 255–262, Sep. 2024, doi: 10.1109/IC3I61595.2024.10829132.

[28] N. Zaidkilani, M. A. Garcia, and D. Puig, “Dual-Stream CoAtNet models for accurate breast ultrasound image segmentation,” Neural Comput Appl, vol. 36, no. 26, pp. 16427–16443, Sep. 2024, doi: 10.1007/S00521-024-09963-W/METRICS.

[29] Y. Liu, Z. Qiu, and X. Qin, “Nested-TNT: Hierarchical Vision Transformers with Multi-Scale Feature Processing,” Apr. 2024, Accessed: Apr. 17, 2025. [Online]. Available: https://arxiv.org/abs/2404.13434v1

[30] A. Sriwastawa and J. A. Arul Jothi, “Vision transformer and its variants for image classification in digital breast cancer histopathology: a comparative study,” Multimed Tools Appl, vol. 83, no. 13, pp. 39731–39753, Apr. 2024, doi: 10.1007/S11042-023-16954-X/METRICS.

[31] T. A. O’Shea-Wheller, A. Corbett, J. L. Osborne, M. Recker, and P. J. Kennedy, “VespAI: a deep learning-based system for the detection of invasive hornets,” Communications Biology 2024 7:1, vol. 7, no. 1, pp. 1–11, Apr. 2024, doi: 10.1038/s42003-024-05979-z.

[32] A. R. W. Sait, “A LeViT–EfficientNet-Based Feature Fusion Technique for Alzheimer’s Disease Diagnosis,” Applied Sciences 2024, Vol. 14, Page 3879, vol. 14, no. 9, p. 3879, Apr. 2024, doi: 10.3390/APP14093879.

[33] N. A. Aljarallah, A. K. Dutta, and A. R. W. Sait, “Image classification-driven speech disorder detection using deep learning technique,” SLAS Technol, vol. 32, p. 100261, Jun. 2025, doi: 10.1016/J.SLAST.2025.100261.

[34] E. Şahin, D. Özdemir, and H. Temurtaş, “Multi-objective optimization of ViT architecture for efficient brain tumor classification,” Biomed Signal Process Control, vol. 91, p. 105938, May 2024, doi: 10.1016/J.BSPC.2023.105938