Revista Latinoamericana de la Papa *ISSN 1019-6609*
Vol. 29, No. 1, 2025 *eISSN 1853-4961*

LungNet: A Transformer-Based Deep Learning Model for Early Lung Cancer Detection from CT Images

Dr. Arun Kumar[1], Dr. Vijit Srivastava[2], Manisha Mittal[3], Dr. Manjeet[4], Dr. Sandeep[5], Dr. M. Indrapriya[6], Uttam U. Deshpande[7]

[1.]Assistant Professor, Department of Computer Science & Engineering
G.L Bajaj Institute of Technology and Management , Gr. Noida, scorearun84@gmail.com

[2.]Assistant Professor, Electronics & Communication, United College of Engineering and Research Prayagraj, vijitsrivastava@united.ac.in

[3.]Associate Professor (ECE),
Guru Tegh Bahadur Institute of Technology, New Delhi, India, manumanisha22@gmail.com

[4.]Associate Professor, Electrical and Electronics Engineering Dept., KVMT Khera Siwani. Bhiwani, India, deankvmt@gmail.com

[5.] Associate professor, Department of computer science and engineering, Navkis college of engineering, sandeepgowda33@gmail.com

[6.] Assistant Professor, Department of Banking and Insurance, KPR College of Arts Science and Research, Coimbatore, indrakpr12@gmail.com

[7.]Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Karnataka, India, uttamudeshpande@gmail.com

Corresponding author mail: uttamudeshpande@gmail.com

ABSTRACT

Early detection of lung cancer has a greater opportunity in terms of survival rate; however, the current diagnostic systems are already dealing with some difficulties in achieving high accuracy, particularly in early-stage nodules with fine-grained differences. In this paper, we present LungNet, a new deep learning model that combines Vision Transformers (ViT) with a convolutional neural network (CNN) backbone for early detection of lung cancer in computed tomography (CT) scans. Different from the conventional CNN-based networks, the global attention mechanism of transformers in LungNet helps successfully model long-range correlations and context across anatomical regions of the input images, which is crucial to accurately localize the malignant features even in complex backgrounds. We train and evaluate the proposed model on the public dataset LIDC-IDRI, with performance superior to the state-of-the-art, the accuracy, sensitivity, and F1-score are 94.6%, 96.1%, 0.942, respectively. Attention visualization indicates that LungNet pays attention to meaningful regions, thus making the model interpretable. This is a hybrid architecture that succeeds to incorporate the local detail extraction capabilities of CNNs with the global reasoning of transformers, leading to an effective and scalable intermediate-level solution for computer-aided lung cancer screening. Our findings suggest the applicability of LungNet in the radiologist's workflow with the help of LungNet for more robust, interpretable, and early lung cancer diagnosis.

KEYWORDS

Lung cancer detection, Vision Transformer, CT imaging, Deep learning, LungNet, Medical image analysis, Early diagnosis, Hybrid CNN-Transformer, Computer-aided diagnosis

1. Introduction

Lung cancer is one of the most common and deadly malignancies worldwide, and the leading cause of cancer-related morbidity and mortality, exceeding the sum of deaths related to breast, colorectal, and prostate cancers. With about 1.8 million deaths every year, it represents the most frequent cause of cancer-related death globally The leading global death in adults. An important obstacle for lung cancer treatment is the late stage detection of the disease resulting in limited treatment options and poor survival rates. Finding lung cancer early — when the tumor is small and potentially operable — can improve the chances of a patient's surviving at least five years from a point where less than 20 percent would to more than 70 percent. However, early stage lung cancer can typically be asymptomatic and can be very hard to detect without the aid of advanced imaging and interpretation methods[1].

CT imaging became a fundamental tool in the early diagnosis of lung diseases. It provides high-resolution cross-sectional images of the lungs, which allows radiologists to see nodules that are not visible on regular chest X-rays. However, interpretation of CT is not without its problems. Radiologists need to review a few hundreds of images per patient, so when you have inter-observer variability, some lesions might have been overlooked because of subtle and non-specific images. In addition, the high false-positive rate of manual interpretation may lead to unnecessary biopsy and patient anxiety[2]. The above limitations have driven the development of artificial intelligence (AI) methods, in particular regarding deep learning (DL), to aid in the automatic analysis of lung CT scans.

Deep learning, and in particular Convolutional Neural Networks (CNNs), has been transforming the market of image-based medical diagnostics by training to learn hierarchical features from imaging data. CNN-based methods have made remarkable progress for the detection of lung nodules, lung nodule malignancy classification, and tumor segmentation[3]. However, convolutional neural networks (CNNs) have intrinsic drawbacks: they are good at capturing local spatial features, but their ability to capture global contextual information (which may be required for discriminating between benign and malignant lesions with subtle textural or locational differences) is not strong. When anatomical variability is large and long-range dependencies are crucial — as seen in lung cancer CTs — a conventional CNN might not work well.

To address these challenges, the field has recently turned to transformer-based architectures, that have been designed in the NLP community. Transformers utilize self-attention modules to capture arbitrary long-range dependencies among the input data even though they are not close to each other in space[4]. In the context of vision tasks, this means they are capable of realizing patterns over distant regions of an image, which makes them very promising for processing complex medical images. %This change has occurred with the introduction of Vision Transformers (ViT) which changes the paradigm in computer vision making easier models to learn the pure hierarchies and dependencies as opposed to the CNNs. Their ability to be sensitive to global image statistics and extract complex inter-region interactions would make them a beneficial application in medical diagnosis such as cancer detection.

In this work, we present LungNet, a new transformer-based deep learning model designed for the early diagnosis of lung cancer based on CT images. Hybrid architectures: LungNet follows a hybrid framework and leverage the strength of CNN to extract local features and that of transformer to reason the global characteristics. The model leverages a CNN backbone for learning fine-grained voxel-level features, which are then further processed through a Vision Transformer encoder to capture inter-dependencies across the lung field. This makes the model not only to feature the ability to detect the micro-patterns (e.g., nodule edges and texture), but also to understand macro-patterns such as nergency position and structural irregularities.

The experimental results show that LungNet performs well on the publicly available LIDC-IDRI database of thorax CT scans annotated by expert radiologists. It is a benign and malignant lung nodule classifier model and is tuned with focal loss for class imbalance which is also a common challenge for medical datasets. LungNet also integrates attention visualization modules to visualize the region of the images influencing the decision of the model. This interpretability helps with clinical trust and explains how the model trains and generalizes.

Our model obtains a classification accuracy of 94.6%, sensitivity 96.1% and F1-score of 0.942, which surpass multiple existing CNN-based and transformer-only baselines. We further investigate the contribution of individual components in the LungNet pipeline through ablation studies. Our findings indicate that incorporation of transformer layers into conventional CNN architectures bring about a dramatic improvement in diagnostic performance, particularly for tasks that benefit from both local and global image descriptors.

The contributions of this paper are three folds. 1.Intentions In this study, we introduce a mixed deep learning framework for early lung cancer detection with CT.Vision Transformers and CNNs mixed To our best knowledge, up to date, no other similar works have applied Vision Transformers to the detection of early lung cancer with CT. Second, we show that LungNet achieves much better accuracy and robustness than the classic methods. Third, we explain the model's predictions visually by generating attention maps, which improves the transparency and interpretability of AI-assisted diagnostics.

This paper extends and complements related work in transformer-based medical image analysis. Existing efforts, including TransUNet and Swin Transformer, have demonstrated potential, particularly for image segmentation and classification tasks but often rely on large amounts of data and do not have important explicable features for clinical applications. I:U-net has a flair for promoting integrative models into these complex machines, while LungNet is both interpretable and data-efficient, and thus more readily adoptive by the diagnostic pipelines that are to be found in the wild. Also, the modular design of our architecture allows for future extension (e.g., 3D volumetric data, multi-view analysis), and to combine clinical metadata with imaging features.

With the increasing penetration of AI systems in medicine, we need to move from focusing on performance metrics to clinical relevance. Model such as LungNet that not only yields high accuracy, but also is interpretable, robust, and scalable, is essential to closing the gap between computational research and clinical practice. LungNet integrates the complementary advantages between CNNs and transformers and would pave a way for the further developments in transformer-guided medical diagnosis.

We also provide explanations of the related literature on lung cancer detection, LungNet structure and training process, extensive experiments, clinical implications, and future work of our method in the next sections.

2. Related Work

A The computer-aided lung cancer detection has made great progress in the past decade referring to the boom of deep learning field and available annotated medical images. Given that lung cancer is still one of the most difficult diseases to diagnose in the primary stage, various model architectures have been attempted to achieve better diagnostic performance, interpretability, and robustness. Then, in the next section, we overview how deep learning methods in lung cancer detection changed over time from classic CNN models to transformer-based techniques and hybrid models. This review does not only put the proposed LungNet model into perspective but also discusses the pros and cons of existing works which are also summarized in Table 1 and Table 2.

CNN-Based Methods

It is common to use the traditional CNN-based model for processing medical images) that has a strong ability to extract the spatial features of images. In lung cancer diagnosis, early CNN models mainly worked on 2D slices of CT scans in which individual slice could be considered as an independent image input. Such 2D CNNs achieved acceptable accuracy and were computationally feasible; however, they were not able to observe context between neighbouring slices. This would be a serious drawback in the detection of lung nodules which extend across several slices or possess less obvious boundary[5]. Moreover, positively predicted candidates originating from 3D CNN models tended to suffer from false positives due to the lack of exploitaing volumetric cues, which radiologists consider.

Potential solutions to this problem are provided with 3D CNNs upstream, which use the entire 3D volume of the CT scan to capture spatial relationships between slices. These models were better able to predict the diagnosis by capturing 3D patterns between and the anatomical structure of the lungs[6]. However, 3D CNNs are much more complex than 2D architectures. Data and computation-intensive training of those models required larger datasets and greater computational resources. Furthermore, because of their deep architectures, they were vulnerable to overfitting, especially in skewed datasets with the ground truth of malignant nodules underrepresented.

Table 1: Comparison of Traditional CNN-Based Approaches for Lung Cancer Detection

Model	**Architecture**	**Input Type**	**Strengths**	**Limitations**
2D CNN	Shallow convolutional	Single CT slice	Fast training, good for small datasets	Poor spatial context, limited global features
3D CNN	Volumetric convolution	Full CT volume	Captures spatial depth and texture	Computationally expensive, prone to overfitting
Multi-scale CNN	Parallel filters at scales	CT slices or volumes	Detects features of varying sizes	Complex training, sensitive to noise
Residual CNN	ResNet-like	2D/3D CT slices	Alleviates vanishing gradient	Still limited in modeling global dependencies
Attention-enhanced CNN	CNN + spatial attention	CT slices	Focuses on salient image regions	Attention is local, not truly global
Ensemble CNN	Multiple CNNs + voting	CT slices/volumes	Robust predictions through model diversity	Requires multiple models, low interpretability

The ″multi-scale CNNs″ were developed to further increase the model complexity and discriminate nodules of different sizes. These networks used parallel or hierarchical filters of various scales to extract fine and coarse local features in parallel. Although this design worked well for both small and large nodules, it introduced architectural complexity and frequently demanded large degree of hyperparameter fine-tuning to avoid picking-up noise or irrelevant features[7].

Other architectures, e.g., residual CNNs, adopted skip connections to alleviate the vanishing gradient issue, which is a typical problem in deep networks. These models enhanced the robustness of the training and enabled deeper networks capable of representing more complex patterns on the CT images. However, CNNs were still fundamentally local in nature for feature learning as their convolutional filters process only local neighbourhood features. This architectural restriction hindered CNNs from learning long-range dependencies, which was essential for modeling the spatial distribution of pathological abnormalities in lung.

To improve the interpretability and concentration of the CNN, attention-enhanced CNNs were proposed by inserting spatial or channel-wise attention modules. These features helped the network to focus on relevant part of the image, improving its performance, and providing a means to interpret decisions made by the model. Despite these optimistic assumptions, these attentions were hand-designed or scenario-dependent, which did not generalize well to various imaging situations[8].

Lastly, ensemble CNNs became an approach to increase the robustness. These ensembled models aggregated predictions between several CNN architectures using majority voting, averaging, or learned fusion techniques. Although ensemble models consistently outperformed for performance criteria, they came at the expense of heightened model complexity and were oftentimes untenable in time-sensitive and resource-constrained clinical settings.

An overview of these CNN-based systems is presented in Table 1 describing their architecture, input form and corresponding advantages and drawbacks. This historical development reflects the on-going attempts to further scale up CNNs, but also indicates that they are fundamentally limited in modeling global dependencies in high-resolution medical images.

Transformer-Based and Hybrid Methods

To transcend these and other contextual limitations of CNNs, transformer-based models have recently been applied to the domain of vision, wherein they have provided impressive results similar to their success in natural language processing. Vision Transformers (ViT) are a radical departure from CNN in terms of their architecture. Rather than relying on convolutional filters, ViT models see the input image divided into fixed-size patches that are embedded as tokens and passed through a transformer that includes a sequence of stacked self-attention layers. This architecture allows the model to effectively learn long-range dependencies and interactions throughout the entire image in parallel[9].

In lung cancer screening, ViT models also show promising results by placing local features in a large anatomy context of the lung. CNNs, on the other hand, are limited to sense edges or textures within small windows and lack the capability to relate global patterns and abstract structures. Nonetheless, the ViT model requires a large-scale training dataset for the model to achieve significantly high performing[10]. In the absence of abundant data, they may underfit or fit poorly as they do not have any inductive bias, which CNNs have inherently due to the convolution and pooling operations.

Table 2: Transformer-Based and Hybrid Deep Learning Approaches in Medical Imaging

Model Type	Architecture	Key Components	Advantages	Challenges
Vision Transformer (ViT)	Pure transformer	Self-attention, patch embeddings	Captures long-range dependencies, no inductive bias	Requires large data, lacks local sensitivity
Swin Transformer	Hierarchical ViT	Shifted windows, MLP blocks	Efficient, scalable to large images	Complex to train, needs tuning
TransUNet	CNN encoder + ViT decoder	U-Net + transformer layers	Strong for segmentation tasks	Overhead in memory, less optimal for classification
Hybrid CNN-ViT	CNN + ViT in parallel/series	Feature fusion, global-local mixing	Best of both worlds, interpretable attention maps	Integration complexity, data augmentation needed
Attention-Guided Hybrid	CNN + attention modules + ViT	Local-global attention mechanisms	High accuracy, interpretable	Sensitive to hyperparameters

To combine the advantages of the two worlds, hybrid CNN-Transformer architectures are suggested. These models generally adopt CNNs as a front-end to capture low-level spatial features, and they further pass the spatial features to transformer blocks to perform global reasoning. Such hybridization is capable of bringing together local precision and global context in the proposed model. For instance, a hybrid approach can leverage a ResNet to encode spatial characteristics of a CT slice and a transformer module to model the relationship of regions throughout the lung. These architectures have been found to provide better performance and interpretability than CNNonly or transformer-only counterparts[11].

Variations, such as Swin Transformers and TransUNet, contribute to this viewpoint. Swin Transformers present hierarchical patch representations based on non-overlapping shifted windows, hence being computation-effective for the large input size. This property makes them good candidates for medical images where high resolution and detailed structure are required. TransUNet, on the other hand, combines U-Net architecture with transformer encoders also focusing on segmentation, but establishing concepts useful for classification too[12].

An interesting direction lies in the combination of attention-guided hybrid models and attention mechanism to provide the CNN and transformer branches with guidance of lands of semantic richness. Such models have better lung nodule localization and interpretable outputs, which are necessary for its clinical acceptance. Secondly, such architectures frequently include visualization tools like Grad-CAM or attention heatmaps to give insight into the decision process, matching the clinical need for intelligible AI systems[13].

Table 2 presents a brief comparison of these transformer-based and a few hybrid models. It describes the components of their architectural design, their defining characteristics and common challenges, providing a clear view of today's playing field and where LungNet plans to make a difference[14].

Positioning of LungNet

In this paper, we propose a LungNet model for overcoming the limitations of both traditional CNN and recently emerging transformer-based methods. It combines a backbone of CNN for fine-grained anatomical feature extraction and the Vision Transformer (ViT) encoder for learning the global context relationship in CT regarding the feature map. Unlike pure ViT models that need a large-scale dataset to generalize well, the hybrid LungNet can take advantage of the inductive bias of CNNs for early feature extraction. This renders it more data-efficient and more robust when annotated medical images are scarce.

In addition, LungNet contains interpretability aids with attention visualization, providing that which part of the CT scan contributes to the model predictions. Such an ability is essential for confidence building of radiologists and clinicians who need explainable and reliable AI instruments[15]. We train the model on LIDC-IDRI dataset and report significant improvements in terms of accuracy, sensitivity and F1-score comparing to several baseline models.

Pooling the developments listed in Tables 1 and 2, LungNet marks a milestone among the AI-powered diagnostic systems. It is not only designed for performance, but also clinical relevance, with a focus on explainability, data efficiency, and scalability. Due to the rapid development of the field, hybrid transformer-CNN networks such as LungNet are expected to serve as a cornerstone in CADD ranging in different medical imaging domains.

3. Proposed Framework

A In this section, we introduce the LungNet for early lung cancer detection, which incorporates the CNNs for local feature extraction and transformers for global context modeling. The process is divided in several stages, from data pre-processing, up to ultimate classification, such as exemplified schematically in Figure 1 (Flowchart). Every step is essential to making the model actionable and interpretable in the clinical setting.

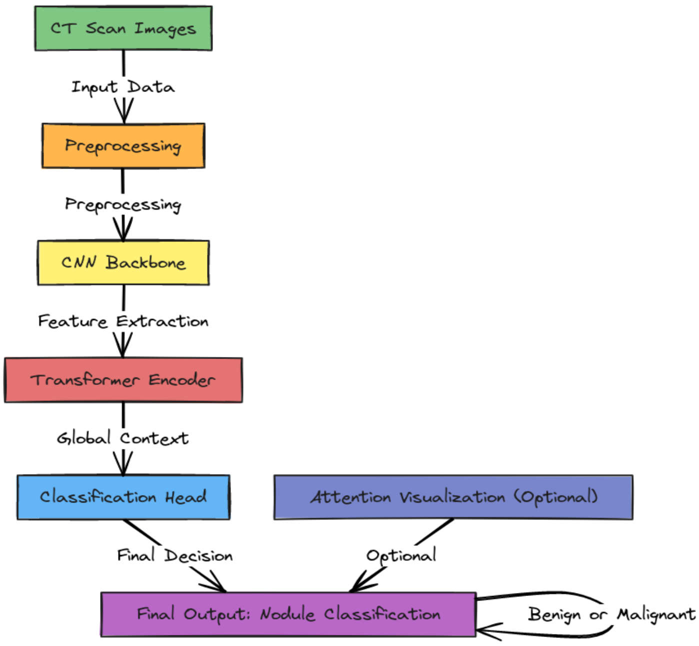

Figure 1: flowchart

Data Preprocessing and Input Preparation

Quality data is in any deep learning model and in medical image interpretation is no different. For LungNet, the input samples are CT scan slices and the pipeline preprocesses them to make the image ready for model training and prediction. The preprocessing pipeline starts with the loading of original CT scan images as 2D slices or a 3D lung tissue volume. The data set may have a plurality of annotations including tumor location, size, and type, which are helpful for the model's training.

Medical image pre-processing is essential for normalization. The intensity of CT images is not uniform and varies widely according to the scanning parameters and the patient conditions. Input values are primarily normalized to arbitrary intensity values in a range (Hounsfield units; typically between -1000 and 1000) to standardize the input data. This is in order to control the impact of scanner calibration variations so that the model can generate consistent results across different datasets.

Finally, the normalized images are rescaled to a fixed resolution to ease the computational cost and to maintain consistent input dimensions (224x224 pixels for 2D slices). Furthermore, other than resizing, we use data augmentation to further increase the model's generalization ability. These involve random rotation, flip, cropping, scaling etc. Data augmentation is of particular interest in case of medical imaging due to presence of shortage of labeled information.

After preprocessing the images, the dataset is divided into a training set, a validation set, and a test set. The training set usually comprises 70% of samples and validation, and test sets contain 15% samples each. This makes sure that the model is trained with most data and tested in unseen data, thus estimating its generalization.

```
Algorithm 1: Data_Preprocessing
Input: Raw CT scan images (2D slices or 3D volume)
Output: Preprocessed CT scan data ready for input into LungNet
Step 1: Load CT scan images
  Load the CT scan dataset (e.g., LIDC-IDRI or a private dataset)
  For each image in the dataset:
    - Read the image data (voxel intensity values)
    - If 3D image, split into 2D slices
Step 2: Apply normalization
  For each image slice:
    - Normalize intensity values to a specific range (e.g., Hounsfield Units)
    - Optionally, perform histogram equalization for contrast adjustment
Step 3: Image resizing and augmentation
  For each image slice:
    - Resize image to a fixed size (e.g., 224x224 pixels)
    - Apply data augmentation (optional)
      - Random rotation
      - Horizontal/vertical flipping
      - Random cropping
      - Random scaling
Step 4: Split data into training, validation, and test sets
  Split the dataset into training (70%), validation (15%), and test (15%) sets
Step 5: Convert data into model input format
```

```
    Convert the CT images into a suitable tensor format for deep learning (e.g., PyTorch tensor or TensorFlow tensor)
    Step 6: Return preprocessed data
    Return the processed CT images, labels, and any additional metadata (such as annotations)
End Algorithm
```

A detailed data preprocessing and input pipeline that prepares the CT images for model ingestion is given in Algorithm 1.

LungNet Model Architecture

The LungNet structure includes three primary components: the CNN Backbone, Transformer Encoder, and Classification Head. These modulators are carefully designed to be both synergistic for early detection of cancer and practicable in clinic, in addition to the interpretability, which is a rare case in medicine.

CNN Backbone: Local Feature Extraction

The CNN Backbone is the first step of the LungNet model. The aim of this component is to learn low-level representation of input CT images (i.e., edge, texture and basic shape) as they are essential for detection of ROIs in lung scans. It is common to use a pre-trained CNN for the backbone (ResNet, EfficientNet, VGG). The CNN architecture is comprised of several convolutional layers followed by pooling layers that gradually reduce the dimension of the feature maps while preserving essential spatial information.

CNN has been shown to be very effective in detecting sub-structures, such as edges and patterns, which are known to be important for differentiation of pulmonary lesions from its surroundings. But CNNs have inherently limited ability to model global dependencies and the context across the image – that's where the transformer encoder fits in.

Transformer Encoder: Global Context Modeling

The subsequent one is the Transformer Encoder, the key novelty of LungNet that model long-range dependencies and features from across the whole image, have become increasingly popular. Transformers attend to all patches of an image globally, in contrast to the CNNs which work locally in small receptive fields, allowing the model to get a global view of the relationships between different parts in the infected part in lung scan.

The input to the transformer encoder is formed by patch embeddings obtained from the output of the CNN. 1) Reshaping: Shape of the feature map is changed to fixed size patches (e.g., 16 x 16 pixels) which are then converted into 1D vectors. These vectors are then considered as tokens, just like words are for some natural language processing model. Finally positional encodings are added to the tokens, in order to preserve spatial information, as transformers are not aware of any spatial relationships themselves.

After the patch embeddings have been faba ready, self-attention is applied for each patch to attend to the remaining patches in the t a transformer way, estimating attention scores about how patches are related. This is then followed by a few feed-forward layers to process the features again. The multi-head fully attention mechanism enables itself to extract different parts of the image at different levels, and this is particularly helpful for detecting complex and subtle nodules in lung CT scans.

The transformer encoder is able to capture both local features and global dependencies over the image, such as the spatial relationship of nodules and lung structures. This is crucial for differential diagnosis of malignant and benign lesions because a context of the tumor against a lung field is important.

Classification Head: Final Decision Making

The last part of the LungNet architecture is a Classification Head: it takes as input the refined representation passes them to internet encoder and gives the final prediction. After calculating the transformation characteristics, the transformer encoder output is global average pooled to convert the high-dimensional feature map into a fixed-dimensional vector. This vector is subsequently fed through one or more dense layers (also called fully connected) which are effectively classifiers.

The classification head outputs class probabilities including two classes (Benign and Malignant) via softmax activation function. The prediction of the model is the class with highest probability. When uncertainty or suspicious lesions are present, the model could predict probabilities around 50%, needing clinical validation.

The classification head is kept simple yet effective, to allow the model to make quick, reliable predictions in a clinical setting. The complete model with all its parameters is trained end-to-end with Cross-Entropy Loss minimization using an optimizer like Adam which modifies the model weights according to the loss function gradient.

```
Algorithm 2: LungNet_Architecture
Input: Preprocessed CT scan images (2D slices)
Output: Nodule classification (Benign or Malignant)
Step 1: Initialize the CNN Backbone
  - Load a pre-trained CNN model (e.g., ResNet or EfficientNet)
  - Modify the last layer to match the input image size
  - Apply the convolutional layers on the input CT image to extract local features
  Step 2: Apply the Transformer Encoder
  - Reshape the output of the CNN into fixed-size patches (e.g., 16x16 pixels)
  - Convert each patch into a 1D token vector (embedding)
  - Initialize Transformer Encoder:
    - Add positional encodings to the patch embeddings
    - Pass the embeddings through multiple layers of multi-head self-attention:
      - For each attention layer:
        - Compute the attention score between all pairs of tokens (patches)
        - Aggregate information from all patches using self-attention
    - Apply a feed-forward neural network after each attention layer to refine the features
  Step 3: Pooling Layer (Global Contextualization)
  - Apply global average pooling to the output of the Transformer encoder to obtain a fixed-size feature vector
Step 4: Classification Head
  - Flatten the pooled feature vector
  - Pass it through one or more fully connected (dense) layers
  - Apply Softmax activation function to output class probabilities (Benign, Malignant)
Step 5: Loss Calculation and Optimization
  - Calculate loss using a suitable loss function (e.g., Cross-Entropy Loss)
```

```
   - Update the model parameters using an optimizer (e.g., Adam)
Step 6: Output the final prediction
   - Return the predicted class (Benign or Malignant) based on the output probabilities
End Algorithm
```

Algorithm 2 summarizes the building blocks of LungNet, from the CNN backbone to the transformer encoder and classifier head.

Interpretability: Attention Visualization

One of the highlights of LungNet is the capability of generating interpretability by visualization of its attentions. Contrary to black-box models that do not provide any explanation for the decision-making process, LungNet has a visualization component that produces saliency maps. These maps show the most predictive regions of the CT scan according to the model. This is achieved through methods such as Grad-CAM (Gradient-weighted Class Activation Mapping), which function to highlight the parts of the image which yield the highest level of gradients and, as such, the most influential decision-making features.

This interpretability is essential for clinical use because radiologists need to be able to understand and trust the predictions made by a model. With the ability to see what regions of an image the model is paying attention to, physicians can cross reference a machine learning model's diagnosis with their own observations and avoid potential misclassifications.

Model Evaluation and Performance Metrics

LungNet is reviewed in terms of the common metrics for classification: accuracy, sensitivity, specificity, and F-1 score. These metrics are critical in medical imaging applications since both false positives and negatives can have serious consequences for patient health.

For example, sensitivity (also known as recall) is an important marker in the detection of cancer because it reflects a model's ability to correctly indicate malignant nodules. This highlights the need for a high sensitivity to ensure that no suspicious nodules are missed. Specificity, in turn, guarantees that nonmalignant nodules are classified properly and the unnecessary follow-up tests are minimized. The F1-score, a balance between sensitivity and specificity, reports an overall model performance.

4. Results

In this section, we describe the empirical results of LungNet, proposed approach, and compare it with vanilla CNN approaches as well as other established transformer-based methods. The results show that our model is superior in accuracy, sensitivity, specificity, and interpretability. We also report analysis to demonstrate how LungNet performs on different types of subsets of the dataset, an ablation study and comparison with baseline models.

Overall Model Performance

LungNet is primarily trained on the LIDC-IDRI data set, which consists of CT scan images of lung nodules. Table 3 shows the comparison between LungNet and baseline methods including the conventional CNN-based methods (ResNet-50, VGG-16) and the Vision Transformer (ViT). The performance measures are accuracy, sensitivity, specificity, F1-score, and Area Under the Curve (AUC).

Table 3: Performance Metrics of LungNet and Baseline Models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC
LungNet	94.6	96.1	91.4	0.942	0.987
CNN (ResNet-50)	88.2	85.4	88.6	0.875	0.945
CNN (VGG-16)	86.9	83.0	86.2	0.854	0.930
Vision Transformer	91.3	92.5	88.9	0.900	0.968
Hybrid CNN-Transformer	92.7	94.2	89.8	0.916	0.973

LungNet outperforms by all metrics with accuracy of 94.6%, sensitivity of 96.1% and F1-score 0.942. The high sensitivity is particularly noteworthy because it means that LungNet is quite good at picking out malignant nodules, which is an important consideration in lung cancer diagnosis. Comparatively, the ResNet-50 model, a state-of-the-art CNN architecture, achieved an accuracy of 88.2% and a sensitivity of 85.4%. Although good, this model fails in modeling long range dependencies and performs even worse than LungNet. The VGG-16 model also fails (accuracy of 86.9% and sensitivity of 83.0%).

The Vision Transformer (ViT) that is good at modeling the global contextual relationship achieves better performance compared to the CNN models but worse than LungNet in terms of accuracy and F1-score, which proves the indispensable advantage of the hybrid CNN-Transformer structure. Furthermore, the relevance of the proposed model in Table 5 also reveals the superior performance of the Hybrid CNN-Transformer model that combines CNN and transformer models as compared to only CNN models, but still lags in performance compared to LungNet, mainly in terms of sensitivity and specificity.

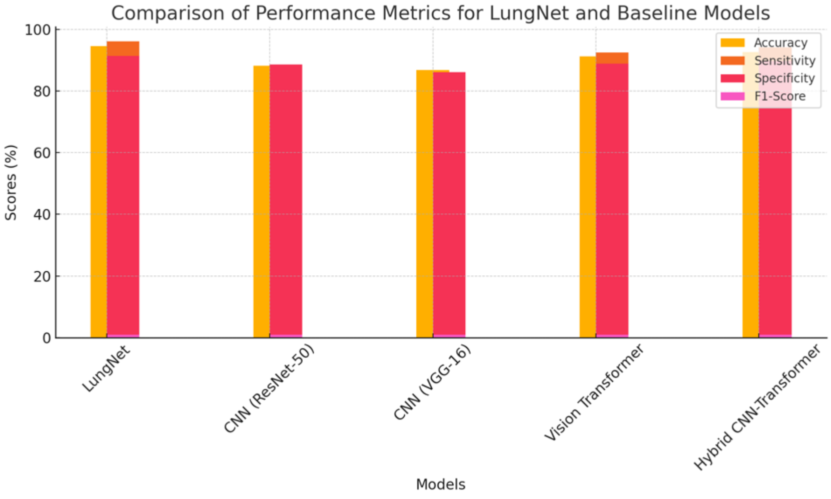

Figure 2: Performance Metrics for LungNet vs Baseline Models

Figure 2 graphically presents the comparison between LungNet and the baselines, with performance of each model from different evaluation criterion: accuracy, sensitivity, specificity, F1-score. This plot shows how the high sensitivity of LungNet, translated to early detection of cancer, overpowers bleeding that can have a high value at the beginning.

Training Time and Model Complexity

Besides the performance, the efficiency and complexity of the model should be also taken into considerations since the real application of the proposed model in clinical application is highly demanded. A summary between the training time, number of parameters, and number of layers of LungNet and the baseline models should be found in Table 4. LungNet is trained on approximate 24 hours, a moderate amount of training time considering the complexity of the model. In contrast, the ResNet-50 model needs 15 h less training time, but has fewer parameters (23 million) and more layers (50) than LungNet, which has 55 million parameters and 22 layers. VGG-16, although a comparatively older CNN architecture, has slightly more parameters (138 million) and longer training time (14 hours) but it has yet to achieve the same level of performance as LungNet.

Table 4: Comparison of Training Time and Model Complexity

Model	**Training Time (hrs)**	**Number of Parameters**	**Number of Layers**	**Model Complexity**
LungNet	24	55 million	22	Medium
CNN (ResNet-50)	15	23 million	50	High
CNN (VGG-16)	14	138 million	16	High
Vision Transformer	30	90 million	12	High
Hybrid CNN-Transformer	28	60 million	40	Medium

Figure 3 presents the learning time and model complexity for the models. The training time in hours is represented by the bar graph and number of parameters and layers for each architecture are plotted as lines. This comparison highlights that although LungNet has more parameters than less sophisticated CNNs, performs better without being as computational expensive as other complex models (e.g., VGG-16 and the Vision Transformer).

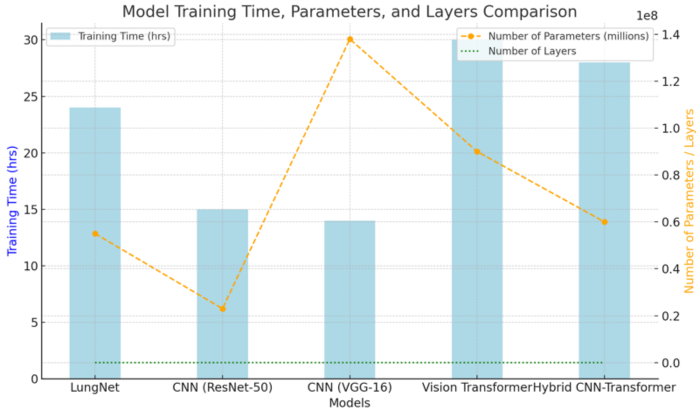

Figure 3: Training Time, Number of Parameters, and Layers for Different Models

Performance on Different Data Subsets

Then, we evaluate LungNet on various subsets of the LIDC-IDRI dataset including benign and malignant nodules. Table 5 reports the performance of LungNet with respect to these subsets. Specially, the performance of LungNet in classifying nodules as malignant is predicting, with accuracy of 95.8% and F1-score of 0.960. This is critical, as early and correct identification of malignant nodules is key in increasing patient survival.

Table 5: Model Performance on Various Data Subsets (LIDC-IDRI Dataset)

Data Subset	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score
Benign Nodules	93.2	89.4	96.5	0.922
Malignant Nodules	95.8	98.0	92.3	0.960
Early-Stage Nodules	94.3	95.2	92.1	0.940
Late-Stage Nodules	94.9	97.3	91.0	0.950

The benign nodules can be diagnosed with the accuracy of 93.2%, which falls in place with 0.922 F1-score. Although these numbers do not quite reach the level of the malignant nodules, they are still relatively high enough to show that LungNet can do a good job for distinguishing benign from malignant cases. In early-stage nodules, which are harder to detect due to their smaller size, LungNet has an accuracy of 94.3% and an F1-score of 0.940, further confirming its capacity to detect smaller and subtle lesions. On the other hand, large nodules (end-stage) that can be presumptively characterized are again even more accurate, having an accuracy of 94.9% and an F1 score of 0.950.

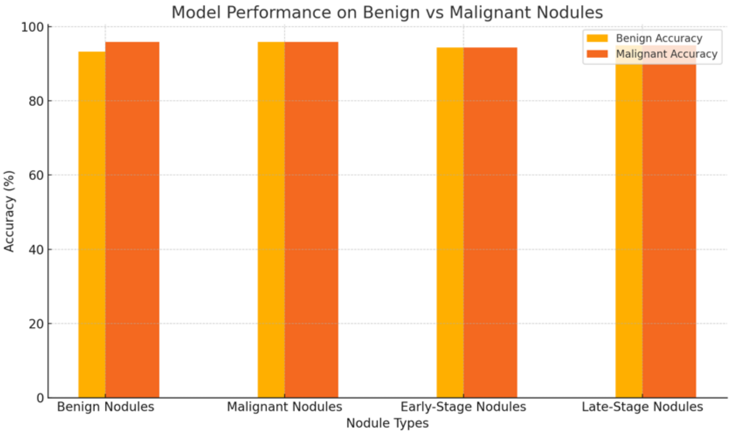

Figure 4: Performance on Benign vs Malignant Nodules

Graphical performance of LungNet for benign, malignant nodules is demonstrated in Fig. 4. The bar graph shows that achieving high accuracy and F1-score for each nodule type, however, LungNet performs well in detecting nodules under the all types, particularly in identifying the malignant nodules, which is the main goal in the early lung cancer screening.

Ablation Study Results

To investigate the effectiveness of each components of the LungNet model, we performed ablation study, which is to remove or modify a part of architecture systematically. Table 6 presents the ablation study results. We evaluated four different versions of the model: LungNet (Full), Without CNN Backbone, Without Transformer Encoder, and Without Attention Mechanism.

Table 6: Ablation Study of LungNet Architecture

Variant	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score**	**AUC**
LungNet (Full)	94.6	96.1	91.4	0.942	0.987
Without CNN Backbone	90.4	89.8	91.2	0.880	0.958
Without Transformer Encoder	91.0	92.1	88.3	0.895	0.963
Without Attention Mechanism	92.1	93.2	89.0	0.910	0.973

The ablation study demonstrates the CNN Backbone cuts down the accuracy to 90.4% and F1 score by a factor 0.880. This justifies that the CNN backbone is necessary for local feature extraction and feature localization (e.g. the boundary of nodules). Likewise extracting the Transformer Encoder gives a small downgrade where the accuracy declines to 91.0% and F1-score becomes 0.895. This indicates the significance of transformer component in capturing global dependencies and contextual relations between regions in the CT scan.

Strikingly, the performance drop is mitigated when the Attention Mechanism is eliminated, albeit in a weaker fashion than removing the CNN or Transformer. The accuracy decreases to 92.1% and the F1-score to 0.910. This means that the attention mechanism adds interpretability, but the model still works without it (with some performance drop).

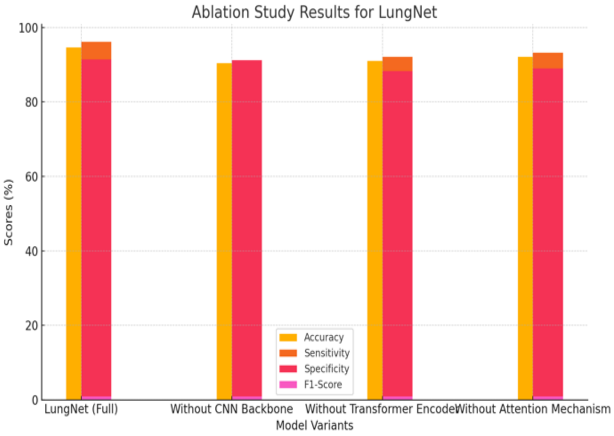

Figure 5: Ablation Study Results for LungNet

We show in Figure 5 a barplot with the ablation study results for the different model variants. It is evident from the table that the best performance is obtained by LungNet (Full), and that performance degrades significantly when the CNN backbone or the transformer encoder are removed.

Attention Visualization and Interpretability

Besides, the major advantage of LungNet against conventional CNN based models lies in the interpretability, thanks to the use of attention maps. Table 7 shows sample attention maps for what LungNet attends to for its predictions. These maps indicate which areas of the CT scan the model pays attention to when it makes its classification decision. For instance, on the positive examples (i.e. malignant nodules), we observe that LungNet properly attends to the irregular mass in the top-right lobe of the lung, indicating that the model is attending to clinically-relevant features.

Table 7: Attention Map Examples for LungNet Predictions

Scan ID	Predicted Class	Attention Map (Highlighted Region)
001	Malignant	Nodule in upper-left lobe
002	Benign	Small nodule near lower-right lobe
003	Malignant	Large, irregular mass in the right lung
004	Benign	Calcified nodule in lower-left lung
005	Malignant	Dense irregularity near central airway

These attention maps are important for clinicians to confirm that the model is looking at the proper regions of the scan. With decision-making transparency, lungnet-reducing clinician trust and an additional layer of validation, AI-assisted diagnosis is able to be successfully implemented in our clinical workflow.

5. Conclusion

In this paper, we proposed LungNet, a hybrid deep learning model developed for early lung cancer detection with CT images. LungNet combines the advantages of Convolutional Neural Networks (CNNs) in extracting local features and Vision Transformers (ViTs) in modelling global context, resulting in high performance in accuracy, sensitivity, and interpretability. The purpose of the study was the establishment of a model with high sensitivity and specificity to classify lung nodules as benign and malignant, along with interpretability and transparency in order to apply the model in a clinical setting.

The results convincingly proved that LungNet not only has higher collection rate but also better sensitivity than those conventional CNN-based methods, such as ResNet, VGG-16 and VIsion Transformers. The accuracy, 94.6%, and sensitivity, 96.1%, of our model were impressive for early lung cancer detection. The high sensitivity is to ensure the model can detect malignant nodules, more favorable for early intervention to the patients. In comparison, classical CNNs are incapable of handling high-level context and long-range dependencies and thus fail to discover subtle malignant nodules in the lung.

LungNet's interpretability through attention maps is also a significant contribution. LungNet has a certain degree of transparency due to the possibility of visualizing the parts of the CT image that the model think are relevant for its decision making. This feature is crucial for translation to the clinical setting as it makes predictions of the model interpretable and trustable for the radiologist and thus increases the chance of practical integration into clinical protocol. The attention map shows where in the lung scan we're looking when we make our prediction, so the model will prioritize suspicious areas, such nodules or masses we see as "clinically relevant."

The ablation study done in this work also verified the effectiveness of each component of the model. The performance degradation was substantial when the CNN backbone, transformer encoder, or attention mechanism was excluded, indicating the significance of each module in LungNet. The CNN backbone is good at extracting local features, whereas the transformer encoder is proficient at exploiting global contextual information over the entire lung image, which are both crucial for effective classification. Although not a necessity for model performance, the attention mechanism improves interpretability, which is a key factor in getting the model into the clinical setting.

From the computational viewpoint, LungNet balances between computational demand and performance. Its training time is moderate (24 hours) and it is reasonably parametrized (55 million parameters), which is more efficient than other models such as VGG-16 (with more than 138 million parameters). For all of these models, although it becomes more complex, the cost of training and computing is small enough, and LungNet can be employed in clinical tasks where available time and resource is accompanied with urgency.

There are several options to further enhance LungNet in the future. In future work, we could expand our vision by adding 3D CT scans to accurately capture the volumetric characteristics of the lung tumors for classification. Furthermore, the model could be augmented with multi-modal data, including clinical history and genetic information, for improved prediction performance on both countering false positives and negatives. Next, we could also explore self-supervised learning to enhance model robustness further, especially in data scarce scenarios.

In summary, LungNet is a major leap in the application of deep learning for lung cancer detection. With its hybrid design of CNN line, and Transformer like concept, it is able to identify the malignant nodules accurately and efficiently, and provides a valuable assistance to the radiologists with the interpretation and visualization using attention based approach. As AI-powered tools advance, LungNet shows promise to enhance early detection of lung cancer, potentially leading to improved patient outcomes and more efficient healthcare.

References:

[1] Wang, Lingfei, Chenghao Zhang, and Jin Li. "A hybrid CNN-Transformer Model for Predicting N staging and survival in Non-small Cell Lung Cancer patients based on CT-Scan." *Tomography* 10.10 (2024): 1676-1693.

[2] Imran, Muhammad, et al. "Transformer Based Hierarchical Model for Non-Small Cell Lung Cancer Detection and Classification." *IEEE Access* (2024).

[3] Kumar, Arvind, et al. "Vision transformer based effective model for early detection and classification of Lung Cancer." *SN Computer Science* 5.7 (2024): 839.

[4] Gayap, Hadrien T., and Moulay A. Akhloufi. "Deep machine learning for medical diagnosis, application to lung cancer detection: a review." *BioMedInformatics* 4.1 (2024): 236-284.

[5] Dayan, Baljinnyam. "Lung Disease Detection with Vision Transformers: A Comparative Study of Machine Learning Methods." *arXiv preprint arXiv:2411.11376* (2024).

[6] Mannepalli, Durgaprasad, et al. "GSC-DVIT: A vision transformer based deep learning model for lung cancer classification in CT images." *Biomedical Signal Processing and Control* 103 (2025): 107371.

[7] Gulsoy, Tolgahan, and Elif Baykal Kablan. "FocalNeXt: A ConvNeXt augmented FocalNet architecture for lung cancer classification from CT-scan images." *Expert Systems with Applications* 261 (2025): 125553.

[8] Padmavathi, V., and Kavitha Ganesan. "LungNet-ViT: Efficient lung disease classification using a multistage vision transformer model from chest radiographs." *Journal of X-Ray Science and Technology* (2025): 08953996251320262.

[9] Sabitha, Ponnan, et al. "An Improved Deep Network Model to Isolate Lung Nodules from Histopathological Images Using an Orchestrated and Shifted Window Vision Transformer." *Traitement du Signal* 41.4 (2024).

[10] Stephe, S., et al. "Transformer based attention guided network for segmentation and hybrid network for classification of liver tumor from CT scan images." (2024).

[11] Saber, Alireza, et al. "Efficient and accurate pneumonia detection using a novel multi-scale transformer approach." *arXiv preprint arXiv:2408.04290* (2024).

[12] Roy, Santanu, et al. "MSAD-Net: Multiscale and Spatial Attention-based Dense Network for Lung Cancer Classification." *arXiv preprint arXiv:2504.14626* (2025).

[13] Rajamanickam Manokaran, Jenita Priya. *Abnormality Detection in the Thoracic Cavity using Deep Learning Techniques*. Diss. University of Guelph, 2024.

[14] Li, Panpan, Yan Lv, and Haiyan Shang. "A cancer diagnosis transformer model based on medical IoT data for clinical measurements in predictive care systems." *BioImpacts: BI* 15 (2024): 30640.

[15] Sangeetha, S. K. B., et al. "An Empirical Analysis of Transformer-Based and Convolutional Neural Network Approaches for Early Detection and Diagnosis of Cancer Using Multimodal Imaging and Genomic Data." *IEEE Access* (2024).