



RESEARCH ARTICLE

LungFormer-XAI: A Dual-Attention CNN–Transformer Framework for Explainable Lung Cancer Detection and Classification learning

Dhatri Raval, Abhilash Shukla*, Atul Patel, Kalpit Soni, Jaimin N Undavia, Unnati Patel

Abstract

Lung cancer is one of the leading causes of cancer-related mortality worldwide, highlighting the need for accurate and timely diagnosis. The capabilities of the Computer-Aided Diagnosis (CAD) system are significantly enhanced with the development of deep learning, but the existing approaches are not sufficient to capture the local and global contextual information and features related to lung lesions, are limited in terms of interpretability, and only utilize a single imaging modality. This research suggests a new multimodal deep learning framework called LungFormer-XAI to overcome these limitation. It uses DenseNet121, Vision Transformers, multimodal feature fusion, Explainable AI (XAI), and dual attention mechanisms to automatically detect and classify Lung Cancer. This method uses Grad-CAM and SHAP to give visual explanations in an understandable manner. Experiment results show that LungFormer-XAI had an accuracy of 97.2%, which is higher compared to other models such as ResNet50 (92.8%), DenseNet121 (94.3%), and EfficientNet-B0 (95.1%). In addition, the proposed framework had a precision of 96.8%, a recall of 97.5%, specificity of 96.6%, an F1 score of 97.1%, and an AUC-ROC of 0.989. Therefore, the results prove that LungFormer-XAI is an accurate and clinically applicable model for automated lung cancer diagnosis.

Keyword: Lung Cancer Detection , Computed Tomography (CT), Vision Transformer Dual Attention Mechanism, Explainable Artificial Intelligence (XAI), Multimodal Learning

Introduction

Lung cancer is one of the most prevalent and deadly cancers in the world, as a very high proportion of cancer deaths and disease is caused by it. The World Health Organization (WHO) places lung cancer at the forefront of the cancers which cause the most deaths, due to its quick progression and often late diagnosis. The ability to accurately classify and identify lung cancer at early stages can significantly improve the survival rate, aid in timely intervention, and minimize

healthcare costs (Lung Cancer, 2024) (Siegel, Giaquinto, & Jemal, 2024).

The traditional approach to the diagnosis of lung cancer has depended on the radiologists reviewing and reviewing hand chest x-rays, Computed Tomography (CT) scans, and histopathological images. While these approaches have been proven to be clinically successful, the diagnostic procedure may be time consuming, laborious, and may be subject to inter-observer variation. Other factors which can affect diagnosis can be the presence of small pulmonary nodules, low-contrast lesions and slight differences in morphology between either benign or malignant nodules (Ardila, Kiraly, & Bharadwaj, 2019) (Setio, Traverso, & De Bel, 2017).

Recent progress in AI and deep learning has revolutionized many tasks related to identifying and categorizing lung cancer, which has proven to be a promising field for automation. In medical image classification, Convolutional Neural Networks (CNNs) and other deep learning algorithms have proven to be highly successful at inferring hierarchical and discriminative features based on medical images without any human engineered features. The models have the ability to identify complex patterns associated

Smt. Chandaben Mohanbhai Patel Institute of Computer, Charusat, Changa, Gujarat , India

***Corresponding Author:** Abhilash Shukla, Smt. Chandaben Mohanbhai Patel Institute of Computer, Charusat, Changa, Gujarat , India , E-Mail: abhilashshukla.mca@charusat.ac.in

How to cite this article: Raval, D., Shukla, A., Patel, A., Soni, K., Undavia, J.N., Patel, U. (2026). LungFormer-XAI: A Dual-Attention CNN–Transformer Framework for Explainable Lung Cancer Detection and Classification learning. *The Scientific Temper*, 17(6):6453-6463.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.6.14

Source of support: Nil

Conflict of interest: None.

with cancerous nodules and support physicians in a more accurate and precise diagnosis, according to (Huang, Liu, Van Der Maaten, & Weinberger, 2017) (Shen, Zhou, Yang, Yang, & Tian, 2017).

A vast amount of varied training data is crucial to the success of deep learning models. Although these two large-scale datasets, LIDC-IDRI and LUNA16, have enabled the development of lung cancer detection systems, challenges remain with generalisation from data-driven models because of their respectively limited scale and lower spatio-temporal resolution, their underrepresented subpopulation—females—and the computational burden of annotating images, differences in imaging protocol, and the absence of clinical metadata. Furthermore, most of existing studies are based on imaging data only, and there is lack of considering additional clinical information related to the patient that is specific to their individual case. This reduces the accuracy of the diagnosis and usability of the results (Armato, McLennan, & Bidaut, 2011) (Setio, Traverso, & De Bel, 2017).

In order to overcome these obstacles and improve the performance of the model, data augmentation approaches, multimodal data integration, and picture preparation are crucial. Techniques can include lung segmentation, image normalisation, contrast enhancement, rotation, scaling, flipping and intensity transformations, to increase dataset variety and increase the robustness of the models to variations between the images (e.g., in acquisition or patient characteristics) in order to enhance the quality of the image. In addition, associated clinical data like gender, age, smoking history, family history and previous lung disease would be valuable to increase accuracy in the diagnosis.

Although there have been significant advances, creating reliable lung cancer detection systems continues to be challenging. However, there are too many existing deep learning models in use today that are focused on extracting features in the image rather than good use of the global context, multimodal clinical data, model interpretability and local feature learning. Furthermore, the black-box behaviour, high false-positive rates and limited generalisability to different datasets of these deep learning models hinder their wide clinical implementation. Despite the success of CNN-based architectures, such as AlexNet, VGGNet, ResNet, DenseNet, EfficientNet, Vision Transformers, and Graph Neural Networks for pulmonary nodule classification (Debnath et al., 2025) (Gai et al. 2024) (He, Zhang, Ren, & Sun, 2016) (Huang, Liu, Van Der Maaten, & Weinberger, 2017) (Dosovitskiy, Beyer, & Kolesnikov, 2021), there is still a need for a complete framework that can integrate the local feature extraction, global contextual modelling, multimodal data fusion, and explainable decision-making.

LungFormer-XAI is the proposed novel framework in the present research, which is centered on the novel

concept of handling lung cancer detection and classification automatically. The proposed architecture features local feature extraction using DenseNet121, lesion-specific feature enhancement through dual attention mechanisms, global contextual learning using Vision Transformer (ViT) encoders, integration of multimodal information from both the computed tomography (CT) image and clinical data, and transparent clinical decision support systems utilizing Explainable Artificial Intelligence (XAI) techniques. The discriminative local properties are learned by DenseNet121 and the clinically important areas are learned by the channel and spatial attention modules. The Vision Transformer part of the system records contextual linkages and long-range spatial dependencies in CT pictures. Furthermore, fusing multimodalities would improve the accuracy of representations provided from imaging data with clinical data and give a more holistic view of the health condition of the patient. The AI-assisted diagnosis approach is enhanced by incorporating explainability features such as Grad-CAM and SHAP-based visualisation techniques for interpretable predictions (Woo, Park, Lee, & Kweon, 2018) (Selvaraju, Cogswell, & Das, 2017) (Dosovitskiy, Beyer, & Kolesnikov, 2021) (Lundberg & Lee, 2017)

Creating a Computer-Aided Diagnosis (CAD) system that can identify lung cancer early on in a reliable, robust, and clinically interpretable manner is the main goal of this project. LungFormer-XAI consists of a framework of multimodal fusion, explainable AI, attention-guided learning, transformer-based contextual modelling, and advanced deep learning, all encapsulated within a single architecture to achieve goals of improved classification accuracy, reduced false-positive detections, enhanced diagnostic reliability, and treatment planning personalisation. The study's ultimate goal is to support the generated development of intelligent (and understandable) AI-centred healthcare systems to enhance clinical outcomes and decrease mortality due to lung cancer.

Literature review

There has been tremendous progress in the identification and diagnosis of lung cancer using artificial intelligence (AI), especially deep learning, which has arisen as a game-changing tool in medical picture analysis. The availability of more and more CT imaging data and enhancement of deep neural network architectures (Ardila, Kiraly, & Bharadwaj, 2019) (Setio, Traverso, & De Bel, 2017) have made it possible to create Advanced Computer-Aided Diagnosis (CAD) for detecting, segmenting, classifying, and assessing the malignancy of pulmonary nodules.

Lung cancer automatic analysis leads to the use of Convolutional Neural Networks (CNNs), which directly extract discriminative features from raw CT scans without any customisation of feature engineering. The latter deep

learning architectures, such as VGGNet, ResNet, InceptionNet, EfficientNet and DenseNet, have inherited architectures of their predecessors that have demonstrated the efficacy of deep CNNs for picture classification (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; He et al., 2016; Huang et al., 2017). Among these networks, the DenseNet model is distinguished by its dense connectivity, feature re-use, and better propagation of gradients. It has been shown to be useful for detecting discriminative features in pulmonary nodules and to avoid overfitting in lung cancer diagnosis (Huang, Liu, Van Der Maaten, & Weinberger, 2017).

Correctly classifying nodules in the target region is vital to improving the performance of the lung nodule classification system. According to several studies, including those by (Oktay, Schlemper, Le Folgoc, & Lee, 2018) (Zhou, Siddiquee, Tajbakhsh, & Liang, 2018) (Ronneberger et al., 2015; Oktay et al., 2018; Zhou et al., 2018), segmentation architectures like U-Net, Attention U-Net, U-Net++, and V-Net have attained state-of-the-art performance in pulmonary lesion localisation and boundary delineation. The diagnostic reliability is improved with these techniques, and are also favorable for the separation of clinically important areas.

An advance in the area of medical imaging for enhancing the performance of deep learning has emerged in the form of attention processes. To extract information that is critical to diagnosis while suppressing background information irrelevant to the task at hand, the channel attention and spatial attention modules are useful for using neural networks in medical imaging, such as computed tomography (CT) scans (Woo, Park, Lee, & Kweon, 2018). A study has indicated that the attention-guided designs are more responsive to the detection of small pulmonary nodules and other subtle features of the lungs compared to standard CNN designs (Durgam, Panduri, Balaji, Khadidos, & Selvarajan, 2025) (Hammad, et al., 2025).

Self-attention mechanism ushered in medical image analysis by modelling global contextual connections and long-range dependences successfully, a capability that is exactly what Vision Transformers (ViTs) can achieve. Introducing the capability of modelling the global contextual connections and long-range dependences, medical image analysis with Vision Transformers (ViTs) stepped into an era of unprecedented achievements. Transformer architectures have been demonstrated to be effective across a few medical imaging applications, which are able to produce more holistic representations of the picture than traditional CNNs that tend to focus on local features (Dosovitskiy, Beyer, & Kolesnikov, 2021) (Wang et al., 2025). Recently, the hybrid CNN-Transformer model has gained much attention regarding enhancing the diagnostic performance (Hatamizadeh, Tang, Nath, & Yang, 2022). These structures merge the local feature extraction ability of CNNs and the global contextual modelling capability of

transformers.

A crucial research field in the diagnosis of lung cancer is data fusion of multimodal data. Recent studies have demonstrated that using clinical information, such as age, gender, smoking history, family history, and pathological data, can significantly boost classification accuracy and practicality over the conventional imaging-only based CAD systems ((Ali et al., 2023;). Multimodal fusion frameworks are more robust and generalisable than single-modality approaches due to their ability to utilise complementary information from various data sources (Oncu & Ciftci, 2025) (Ayadi, et al., 2025).

AI techniques, in particular generative AI, also offer a stepped response to the challenges of limited training data. Researchers Frid-Adar, Klang, Amitai, Goldberger and Greenspan (2019) increased the resistance of the model to the effects of noise and decreased the risk of overfitting by using Generative Adversarial Networks (GANs) to synthesize pulmonary nodules and augment data (Frid-Adar, Klang, Amitai, Goldberger, & Greenspan, 2019). More recently, Javed et al. (2024) and Ma, Zheng, Chen, Qi, and Rong (2025) reported that diffusion-based-generative models may enhance the image quality and enhance missing data in low-dose CT imaging systems (Javed, Abbas, Khan, Daud, Bukhari, & Alharbey, 2024) (Ma, Zheng, Chen, Qi, & Rong, 2025).

Despite some promising findings regarding the diagnosis of lung cancer, there still remain some challenges to overcome. Few methods integrate the two methods, one for contextual learning and the other for local feature extraction by CNN. Moreover, there is much research only examining imaging data, and failing to benefit from complementary clinical data. Other complications with interpretability, false-positive rates and generalizability across datasets are preventing its widespread clinical applications (Lundberg & Lee, 2017) (Selvaraju, Cogswell, & Das, 2017). These strategies ensure regulatory acceptance of AI assisted diagnostic systems, contribute to increased transparency and clinician confidence.

Thus, there exists a requirement for an elaborative, clinically explainable framework that can employ DenseNet like feature extraction, attention like lesion improving, transformer like contextual learning, multi-modal data fusion and explainable AI together. Given these challenges, this research introduces the state-of-the-art framework of LungFormer-XAI aimed at the detection and classification of lung cancer based on DenseNet, Dual Attention, Vision Transformer, and Explainable AI.

Research Gap

Although there has been great progress in the development of deep learning-enabled CAD systems, the accurate and early detection of lung cancer is still a difficult problem to solve. There have been many studies that have used

CNNs, DenseNet, Vision Transformers, attention models, and multimodal learning; however, very few studies have utilized the power of these complementary technologies in combination with each other. (Ardila, Kiraly, & Bharadwaj, 2019) (Dosovitskiy, Beyer, & Kolesnikov, 2021) (Shen, Zhou, Yang, Yang, & Tian, 2017) (Revathi & Leni, 2025), all propose that this implies issues with imperfect feature representation, subpar generalisation and classification performance remain.

Furthermore, many approaches rely solely on CT images while overlooking important clinical information such as age, sex, smoking history, and family history. The limited adoption of explainable AI also reduces clinicians' trust in automated predictions (Durgam, Panduri, Balaji, Khadidos, & Selvarajan, 2025) (Hammad, et al., 2025) (Goswami, Chandra, & Kar, 2026). Therefore, there remains a need for an integrated framework that combines local feature extraction, attention-guided learning, transformer-based global context modeling, multimodal data fusion, and explainable AI to provide accurate and clinically interpretable lung cancer diagnosis.

The main scientific contribution of this research is the introduction of LungFormer-XAI, which is a unifying framework that combines the following elements in one framework: DenseNet121 for local feature extraction, dual attention mechanism for feature representation improvement, vision transformer for global context learning, multimodal data fusion, taking into account the information about patients' disease, and Explainable Artificial Intelligence methods (Grad-CAM and SHAP) for explainable diagnosis of lung cancer.

Objectives of the Study

Despite it being one of the leading causes of death from cancer worldwide, prompt detection of lung cancer has a big improvement in survival and treatment outcomes. Improvements in explainable AI (XAI), attention processes, and deep learning have the potential to bring more accurate and reliable computer-aided diagnosis (CAD) of lung cancer closer. Therefore, it is necessary to develop a cutting-edge multimodal deep learning framework: LungFormer-XAI. Thus, it is crucially important to devise a state-of-the-art multimodal deep learning framework: LungFormer-XAI, with the purpose of achieving automatic detection and classification of lung cancer from computed tomography (CT) scans. The point (s) of the study are :

- Creating a deep learning framework (LungFormer-XAI) that can be explained and used automatically to accurately identify and classify lung cancer from CT scans.
- In order to improve feature representation and lesion localisation, the Vision Transformer (ViT) architecture, dual attention processes, and DenseNet121 will be combined.

- Multimodal imaging will combine clinical information (age, gender, smoking, family history) with imaging characteristics of CT to increase the accuracy and precision of the diagnosis.
- The model can be made more resilient and accurate for classification using methods such as hyper-parameter tuning, data augmentation, using weighted losses, and transfer learning.
- The idea is to create predictions that can be interpreted by health-care providers via XAI strategies such as Grad-CAM or SHAP. This framework will then be assessed based on various metrics such as Accuracy, Precision, Recall, F1-Score, Sensitivity, Specificity, and AUC-ROC.

Dataset Description

The proposed LungFormer-XAI architecture is evaluated on two lung cancer imaging datasets that are publicly available: LIDC-IDRI and LUNA16 mention in table 2. Therefore these images have been labelled as industry standards for pulmonary nodule identification and classification studies, due to their annotated thoracic CT images, according to Armato, McLennan, and Bidaut (2011) and Setio, Traverso, and De Bel (2017) (Armato, McLennan, & Bidaut, 2011) (Setio, Traverso, & De Bel, 2017).

Our LIDC-IDRI database contains 1,018 thoracic computed tomography scans from several hospitals. The pulmonary nodules and malignancy scores were indicated and provided by up to four radiologists independently, for each imaging. This data set can be used for machine learning training to detect lung cancer as the nodules are of different sizes, shapes and malignancy levels.

The LUNA16 dataset is derived from LIDC-IDRI with 888 CT images, with inconsistent annotations or slice thickness > 3 mm, removed mention in table 3. It has been extensively used to assess automated systems for detecting lung nodules, but provides standardized characterization of nodules larger than 3 mm.

The pre-processing steps carried out on all the CT images prior to training involve image scaling, intensity normalization, contrast enhancement and noise reduction. In addition to this, lung segmentation is performed as a part of the pre-processing. Data augmentation techniques such as flipping, rotating, scaling, translating and randomly varying intensities are implemented to increase the generalisability of the model and reduce overfitting. For objective evaluation of the performance, the data set is divided into three sets: training (70%), validation (15%) and testing (15%).

Model Implementation and Training Configuration

An NVIDIA GPU-enabled computing environment is used to train and infer models faster in the proposed LungFormer-XAI framework, which is based on the TensorFlow and PyTorch deep learning frameworks. In this system, lung

Table 1: Research Gap Analysis of Existing Lung Cancer Detection Approaches

Study	Methodology	Strengths	Limitations	Research Gap
Javed et al. (2024)	Comprehensive review of deep learning techniques for lung cancer detection	Summarizes recent AI advancements and diagnostic trends	Primarily review-based; no unified implementation framework	Lack of integrated multimodal and explainable AI framework
Faizi et al. (2025)	Deep learning-based CT image classification	High classification performance on CT images	Focuses mainly on imaging features	Does not integrate clinical metadata and explainability
Abdulqader et al. (2025)	Multi-objective deep learning for tumour detection, localization, and classification	Simultaneously improves multiple diagnostic tasks	Computationally complex and imaging-centric	Limited multimodal fusion and interpretable decision-making
Durgam et al. (2025)	Hybrid CNN-Transformer architecture for lung cancer detection	Combines local and global feature learning	Explainability mechanisms not incorporated	Lack of clinically interpretable predictions
Oncu & Ciftci (2025)	Multimodal AI framework integrating CNN and ANN with imaging and clinical data	Improves diagnostic accuracy using heterogeneous data sources	Limited attention-based feature refinement	Absence of transformer-based contextual modelling and XAI
Hammad et al. (2025)	Custom CNN with Explainable AI for lung cancer diagnosis	Provides interpretable diagnostic outputs	Relies on CNN-only feature extraction	Does not leverage transformer architectures and multimodal fusion
Nithya & Vidhyapathi (2025)	Transfer learning-based deep architecture for CT image classification	Efficient feature extraction and improved classification accuracy	Uses imaging modality only	Limited utilization of clinical information and explainable AI
Zhou et al. (2025)	Ensemble deep learning models for CT-based lung cancer classification	Enhanced robustness and classification performance	Increased computational complexity	Lack of integrated attention mechanisms and interpretability
Goswami et al. (2026) , Haque et al. (2025)	Multi-model deep learning framework with Explainable AI	Combines multiple deep learning models with XAI	Limited multimodal data integration	Need for unified CNN–Transformer–Multimodal architecture
Proposed LungFormer-XAI	DenseNet121 + Dual Attention + Vision Transformer + Multimodal Fusion + Grad-CAM + SHAP	Combines local and global feature learning, multimodal fusion, explainability, and clinical data integration	Computational cost associated with transformer-based architectures	Addresses limitations of existing methods through an interpretable multimodal framework

cancer is automatically identified and classified with the help of Computed Tomography (CT) images. It also introduces DenseNet121-based feature extraction, dual attention methods, Vision Transformer (ViT) encoders, multimodal feature fusion and Explainable Artificial Intelligence (XAI) modules.

After being trained using ImageNet pre-trained weights, the DenseNet121 backbone is transferred-learned-to-be fine-tuned on the lung cancer dataset. In order to be fed into the network, the CT scans undergo a number of preliminary steps, such as separating the lungs, scaling down the picture to 224×224 pixels, normalising intensity and enhancing contrast. The learning rate is dynamically adjusted from the initial learning rate value 1×10^{-4} , using a Reduce-on-Plateau scheduler that has dependency on the validation loss. During training, the Adam optimiser is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to guarantee efficient and stable convergence.

In order to achieve a happy medium between computational performance and GPU memory utilisation,

a batch size of 32 was chosen. To avoid overfitting, the model is trained for 50 epochs and has an early stopping mechanism that kicks in when validation loss is unchanged for 10 consecutive epochs. To provide network stability, batch normalisation (BN) layers are added throughout the network. Fully connected layers are dropouts with a dropout rate of .5 to improve the generalisation performance.

A set of data augmentation methods are used in training, such as random rotation, horizontal flip, scaling, translation, zoom and intensity adjustment, to enhance training resilience while reducing overfitting. These augmentation strategies may boost the generalisability of the model to other CT imaging collection methods and patient characteristics.

The dual attention module can use channel attention and spatial attention techniques to isolate specific data regarding the attention that has been set to be paid to the diagnostic part of the lesion, while allowing data that may contain “noise” to be ignored. DenseNet121 subsequently passes its feature representations to the Vision

Table 2: Dataset Characteristics

<i>Dataset</i>	<i>Image type</i>	<i>Number of scans</i>	<i>Annotation type</i>	<i>Purpose</i>
IIDC-IDRI (Armato, McLennan, & Bidaut, 2011)	Thoracic CT Images	1,018 CT scans	Nodule location, contour, malignancy rating	Lung nodule detection and classification
LUNA16 (Setio, Traverso, & De Bel, 2017)	CT Images (subset of IIDC-IDRI)	888 CT scans	Pulmonary nodules (>3 mm)	Benchmark evaluation
Combined Dataset	CT Images	1,906 scans	Benign and malignant nodule annotations	Training and testing LungFormer-XAI

Transformer encoder, to extract long-range dependencies and global structural connections from the CT image. The use of a multimodal fusion layer, combining the imaging characteristics retrieved with clinical parameters such as gender, age, smoking history and family history, is used to boost classification ability.

To optimize, the network considers the imbalance of the classes of benign and malignant nodules and optimizes using a weighted categorical cross-entropy loss function. A gradient of 1.0 is used for gradient clipping to ensure a stable training while preventing the gradients from explosion. Furthermore, a warm-up method with learning-rate is applied on some of the initial training epochs to facilitate the convergence of the components of the transformer.

Hyperparameters like learning rate, batch size, dropout rate, attention weights, and transformer embedding dimension are fine-tuned with the assistance of validation set. Final model configuration is obtained by selecting the appropriate configuration based on classification performance indicator including SNR, AUC-ROC, accuracy, precision, recall, F1-Score and specificity and sensitivity.

The end configuration of the LungFormer-XAI model was determined via hyper-parameter tuning based on the validation. The picture size of the input was 224×224 pixels and DenseNet121 was deployed as the backbone network. The Adam optimiser gradient was trained in the model with values: initial learning rate of 1×10^{-4} , 32 batches, learning rate decay of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. After 10 completed iterations of training, we stopped training if we observed no improvement in the validation set. We made a total of 50 iterations, holding off after every 10 episodes for any improvements made in validation. A Reduce-on-Plateau learning-rate scheduler, a weighted categorical cross-entropy loss, batch normalisation, a dropout rate of 0.5 and gradient clipping (value equal to 1.0) were used to increase stability and generalisability of training. For interpretability, we applied Grad-CAM and SHAP to determine model interpretability, and for performance metrics, we used Accuracy, Precision, Recall, F1-Score, Sensitivity, Specificity and AUC-ROC.

Proposed LungFormer-XAI Framework

The proposed LungFormer-XAI model follows a sequential process flow for diagnosing and classifying lung cancer. First,

Table 3: Types of Images Available in IIDC-IDRI

<i>Image category</i>	<i>Description</i>
Normal Lung CT	CT scans without clinically significant nodules
Small Pulmonary Nodules	Nodules less than 3 mm diameter
Benign Pulmonary Nodules	Non-cancerous lung lesions
Malignant Pulmonary Nodules	Cancerous lung lesions
Non-Nodule Findings	Scars, calcifications, and other abnormalities

the CT images are preprocessed through normalization, resizing, and augmentation, along with the collection of clinical data. Discriminative local features are then captured through DenseNet121, and a dual attention mechanism is used to emphasize the regions of diagnostic interest. Then, global contextual relations are captured through a vision transformer, and the local features are fused with clinical data through a multimodal fusion approach. Lastly, the fused features are classified, and Grad-CAM and SHAP approaches are used for explainability.

The initial noise removal and shrinking of images, followed by lung segmentation, are the first effective steps in CT processing that can be taken when activity intensity normalisation is not completed, when intensity normalisation is not done, noise removal and shrinkage of the images is the first important steps in processing CT images, followed by lungs segmentation. The techniques used in the pre-processing steps are helping to improve the quality of the images and to easily extract useful features from pulmonary nodules. To promote model generalisation and reduce overfitting, techniques such as data scaling, rotation, horizontal flipping, translation and intensity modulation are employed in the training of the model.

DenseNet121 is employed in the step of feature extraction. The dense connection structure and the effectiveness of its feature reusability led to its good performance in medical picture classification (Huang, Liu, Van Der Maaten, & Weinberger, 2017). Although many deep learning models experience the vanishing gradient phenomenon, DenseNet121 is able to overcome it and capture a hierarchy of pulmonary nodule representations. The remarking of feature maps obtained is sent to a dual

attention module with both channel and spatial attention mechanisms. During this module, the network focuses more on the diagnostically important regions instead of irrelevant background information, as lesion localisation improves and feature representation is enhanced (Woo, Park, Lee, & Kweon, 2018).

The optimized feature maps are then fed into a Vision Transformer (ViT) encoder to obtain the mapping result. Unlike conventional CNNs that typically focus on local features, (Mannepalli et al. (2025)) (Dosovitskiy, Beyer, & Kolesnikov, 2021), claim that vision transformers are capable to model relationships across long-range and global context in CT images using their self-attention processes. This allows the framework to better identify the structures that make up the somewhat intricate characteristics of malignant lung nodules.

A multimodal fusion layer could improve the diagnostic reliability by combining clinical (sex, age, smoking and/or familial history) with imaging information that is extracted from the CT scan. The fusion process, which combines a multitude of complementary information sources, produces a complete feature representation for classification. The final layer of classification is based on a weighted categorical cross-entropy loss, which is used to address class imbalance and make a diagnosis of either benign or cancerous for a lung nodule.

As part of the proposed strategy, one of these methods is Explainable Artificial Intelligence (XAI), a tool developed to increase the openness and confidence of the clinicians. Grad-CAM generates heatmaps indicating the relevance of different regions in lungs for predictions, and SHAP measures the contribution of individual imaging and clinical parameters in diagnosis (Selvaraju, Cogswell, & Das, 2017). The outputs that these explainability processes provide are interpretable and make the framework more useful in healthcare applications and improve clinical decision making.

Developing the Suggested LungFormer-XAI System using Mathematical Means

Let the set of lung x-rays, say, be denoted as.

$$D = \{(X_i, Y_i)\} \text{ where } i = 1 \text{ to } N$$

in, where the output is the class label Y_i corresponding to the given CT image X_i from set $\{1, N\}$.

Image Pre-processing

$$X_p = T(X_i)$$

In this case this amounts to the pre-processed picture X_p and the segmentation, resizing, normalisation and augmentation procedures T .

DenseNet121 Feature Extraction

The pre-processed picture is fed into the DenseNet121 backbone which extracts deep features. The representation of the retrieved features is taken as:

$$F_D = \phi_D(X_p)$$

The DenseNet121 feature extraction function is ϕ_D represented by the FD Deep feature map that is extracted by it.

Channel Attention

Highlighting areas of the lungs which are diagnostically important requires a two-pronged approach, both channel and spatial. one can obtain the feature representation using attention enhancement, through

$$A_c = \sigma \left[W_2 \left(\text{ReLU} \left(W_1 (Z_c) \right) \right) \right]$$

Refined feature map

$$F_C = A_c \times F_D$$

Where:

- A_c = Channel attention weights
- σ = Sigmoid activation function

Spatial Attention

$$A_s = \sigma \left[\text{Conv} (F_{avg}, F_{max}) \right]$$

Refined spatial features

$$F_a = A_s \times F_C$$

Where F_{avg} is Average pooled features and F_{max} is Maximum pooled features

Vision Transformer Encoding

$$F_t = \phi_D ViT(F_a)$$

Where F_t is Transformer feature representation

Clinical Metadata Encoding

$$M = \{m_1, m_2, m_3, \dots, m_K\}$$

$$F_m = \phi_m(M)$$

Where F_m is Clinical metadata features

Multimodal Fusion

$$F_{fusion} = F_t \oplus F_m$$

Where $\oplus$ denotes feature concatenation

Classification Function

$$P(Y|X) = \text{Softmax}(WF_{fusion} + b)$$

Where W is Weight matrix and b is Bias vector

$$\hat{Y} = \arg \max [P(Y|X)]$$

Where $\hat{Y}$ is Predicted class, Benign = 0 and Malignant = 1

$$L = -\sum [w_i Y_i \log(\hat{Y}_i)]$$

Where: L is Total loss, w_i is Class weight, Y_i is True label and $\hat{Y}_i$ is Predicted probability after that apply adam optimizer.

Overall LungFormer-XAI Framework

$$\hat{Y} = f_{XAI} [\phi_{ViT} (\phi_{CA} (\phi_{SA} (\phi_D (T(X)))) \oplus \phi_M(M)]$$

This equation represents the complete workflow.

The simplified LungFormer-XAI process is presented in Figure 1. It starts with a CT Scan Input that goes via deep Feature Extraction (DenseNet). Then, it comes to a refinement step where it captures both local and global contextual data using a Dual Attention & ViT block. Data flow to the Feature Fusion step passes into a XAI Explainability engine which outputs a Combined Output in the end. An XAI Insights visualisation of the Region of Interest and a set of explainability routes that highlight important clinical

indicators like Nodule Size and Smoking History are produced by this explainability module. A final clinical DIAGNOSIS—whether benign or malignant of the nodule that is present in the lungs—is then generated by combining the feature representations and the explanatory insights.

Experimental Results and Discussion

To evaluate the effectiveness of the proposed LungFormer-XAI framework, the quantitative and qualitative evaluations were carried out on the below-mentioned datasets comprising CT images of lung cancer. The framework includes DenseNet121-based feature extraction, dual-attention techniques, Vision Transformer encoding, multimodal fusion, and Explainable Artificial Intelligence (XAI) modules, which boost diagnostic performance and interpretability in clinical applications.

Quantitative Analysis

In a quantitative evaluation the following measures were employed: accuracy, precision, recall (sensitivity), F1-Score, specificity, and ROC-Area under the curve. Outperforming traditional CNN-based methods like ResNet50 (92.8%), DenseNet121 (94.3%), and EfficientNet-B0 (95.1%), the suggested LungFormer-XAI model had an overall classification accuracy of 97.2%.

Additionally, the proposed framework also reduced false-positive and false-negative predictions and achieved a high rate of specificity at 96.6%, F1-Score at 97.1%, Recall (Sensitivity) at 97.5% and Precision at 96.8% for the identification of benign and malignant pulmonary nodules. Furthermore, the model achieved an AUC-ROC value of 0.989 indicating excellent cancer-noncancer power in terms of discrimination

The comparative analysis presented in Figure 2 demonstrates that the proposed LungFormer-XAI framework achieves the highest classification accuracy of 97.2% compared with conventional CNN-based and transformer-based architectures. The improvement in performance indicates that the integration of DenseNet121 feature extraction, dual attention mechanisms, Vision Transformer encoding, multimodal fusion, and Explainable Artificial Intelligence (XAI) significantly enhances feature representation and diagnostic decision-making for automated lung cancer detection.

The findings of the experiment clearly show that the presented LungFormer-XAI approach surpasses traditional CNN models in performance due to the successful combination of local feature extraction, contextual information and multimodal feature integration. High accuracy, recall, and AUC-ROC values reflect the ability to reliably detect both benign and malignant nodules while keeping specificity high enough to reduce false positives. What is more, the use of Grad-CAM and SHAP makes it possible to increase the interpretability of the results.

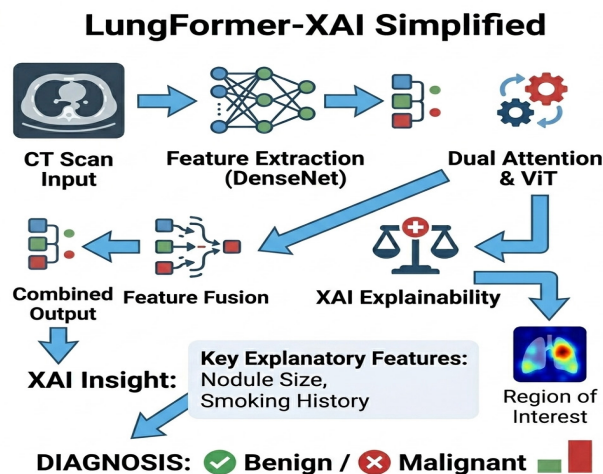


Figure 1: Lung Former-XAI

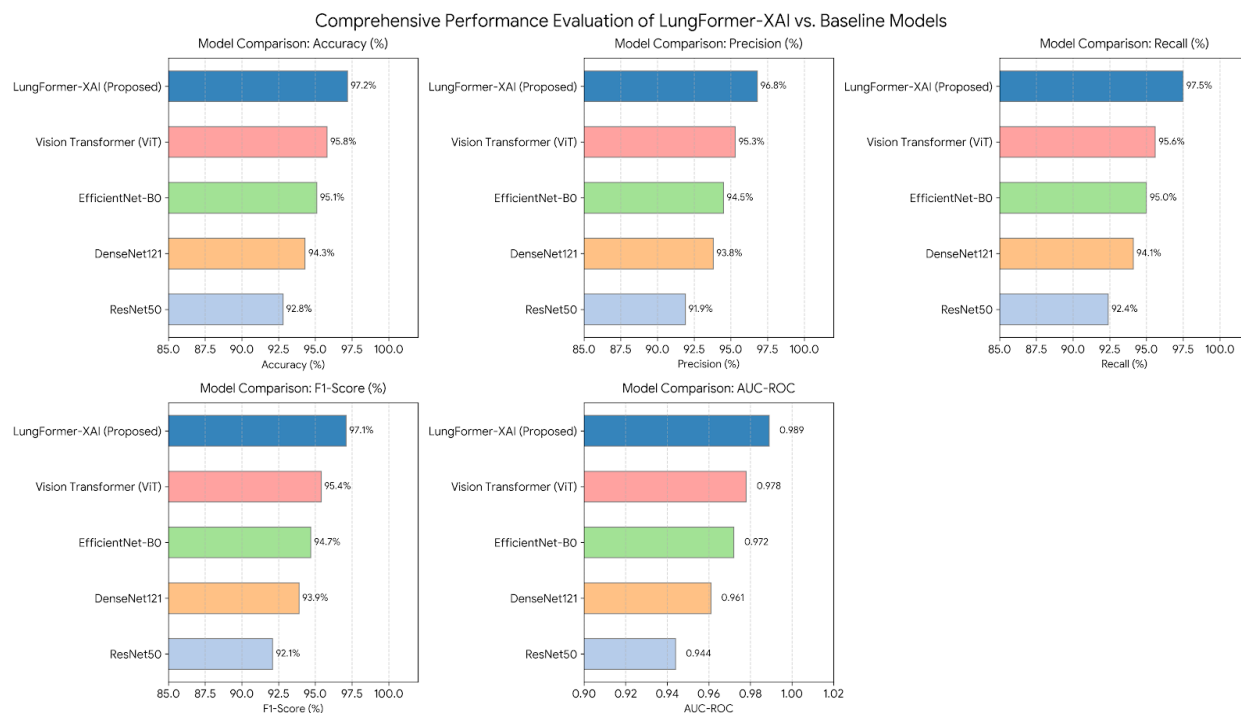


Figure 2: Comparison of Different Models

Comparative Advantages of the Proposed LungFormer-XAI Framework

- Some clinical data such as age, gender, smoking history and family medical history are incorporated with imaging parameters extracted from CT scan.
- DenseNet121 makes good use of dense feature reuse to accurately record morphological features and local texture patterns of lung nodules.
- Channel focus together with spatial attention is applied, thereby making diagnostically important areas of the lungs stand out and irrelevant background data are suppressed.
- With the help of Vision Transformer (ViT), CT images may have their global structural linkages and long-range contextual dependencies captured.
- Multimodal fusion increases diagnostic accuracy and categorisation performance using supplementary clinical and imaging information.
- Weighted categorical cross-entropy loss accounts for the fact that there is a significant class imbalance between benign and malignant nodules.
- Grad-CAM and SHAP visual explanations, as well as feature resolutions provide enhanced diagnostic transparency and clinician confidence.
- This helps to avoid false positive and false negative rates due to the learning of contextual features and improvement of lesion localisation.
- Shows very high generalisability for a variety of CT imaging disorders and patient populations.

- Offers a scalable CAD system for early screening and detection of lung cancer that can be clinically interpreted.

To aid radiologists in the early diagnosis and categorisation of lung cancer, the suggested LungFormer-XAI framework showcases the promise of state-of-the-art deep learning methods. The technique integrates multi-modal data fusion, attention-based learning, transformer contextual modelling, and Convolutional feature extraction for developing comprehensive models of pulmonary nodules extracted from computed tomography (CT) scans. Model transparency is enhanced with the addition of Explainable AI, which identifies the areas and clinical considerations that play a role in diagnostic choices. The suggested framework is a trustworthy and useful tool for helping doctors with lung cancer screening programs because of these features. As deep learning technology continues to improve, interpretable and intelligent systems such as LungFormer-XAI can transform the process of diagnosing lung cancer, improve the efficiency of medical care and provide an opportunity for timely interventions for patients.

Conclusion

An innovative deep learning framework for automated identification and classification of lung cancer using Computed Tomography (CT) images was introduced in this research called LungFormer-XAI. The proposed system combines the principles of Explainable Artificial Intelligence (XAI), multimodal information fusion of imaging and clinical

data, dual attention processes, Vision Transformer (ViT) encoding, and DenseNet121 feature extraction to enhance the system's diagnostic performance and interpretability. It combines the global and local feature learning and is used to improve the classification accuracy and consider complex lung nodule features efficiently.

The accuracy, precision, recall, F1-score, specificity, and AUC-ROC values show that LungFormer-XAI outperforms the traditional CNN-based and transformer-based models. Attention mechanisms aid in lesion localization and multimodal ensembling with additional clinical information enhances diagnostic certainty. The addition of Grad-CAM, SHAP makes more compelling and understandable predictions for enhanced visibility and increased clinical confidence among better predictions from physicians.

Finally, the LungFormer-XAI framework is an accurate, robust and explainable solution for the diagnosis of lung cancer. The concept seems to have tremendous potential for enhancing clinical care through rapid clinical interventions, supporting radiology services with timely diagnosis, and reducing misdiagnosis.

Acknowledgement

The authors acknowledge the institutional support of the faculty of computer applications, Charotar University of Science and Technology, in the preparation of this research.

Declaration of Conflicts of Interest

The authors declare that there are no financial, personal, or professional conflicts of interest that could have influenced the research findings or conclusions presented in this article. The research was conducted independently, without funding from any commercial or institutional body with a direct interest in its outcomes. All authors have read and approved the final manuscript and confirm that the work is original and has not been submitted elsewhere for publication

References:

- Abdulqader, A., Abdulameer, S., Bishoyi, A., Yadav, A., Rekha, M., & Kundlas, M. (2025). Multi-objective deep learning for lung cancer detection in CT images: Enhancements in tumor classification, localization, and diagnostic efficiency. *Discover Oncology*, 16, 529. doi:10.1007/s12672-025-02314-8
- Ali, H., Mohsen, F., & Shah, Z. (2023). Improving diagnosis and prognosis of lung cancer using vision transformers: A scoping review (arXiv:2309.02783). *arXiv*. <https://doi.org/10.48550/arXiv.2309.02783>
- Ardila, D., Kiraly, A., & Bharadwaj, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6), 954-961.
- Armato, S., McLennan, G., & Bidaut, L. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2), 915-931.
- Ayadi, W., Farhat, Y., Althabahi, S., Baskaran, N., Dar, S., Indhumathi, R., ... Rather, A. (2025). AI-Powered CNN Model for Automated Lung Cancer Diagnosis in Medical Imaging. *International Journal of Statistics in Medical Research*.
- Debnath, J., Pranta, A. S. U. K., Hossain, A., Sakib, A., Rahman, H., Haque, R., Ahmed, M. R., Reza, A. W., Swapno, S. M. M. R., & Appaji, A. (2025). LMVT: A hybrid vision transformer with attention mechanisms for efficient and explainable lung cancer diagnosis. *Informatics in Medicine Unlocked*, 57, 101669. <https://doi.org/10.1016/j.imu.2025.101669>
- Dosovitskiy, A., Beyer, L., & Kolesnikov, A. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.
- Durgam, R., Panduri, B., Balaji, V., Khadidos, A., & Selvarajan, S. (2025). Enhancing lung cancer detection through integrated deep learning and transformer models. *Scientific Reports*, 15, 15614. doi:10.1038/s41598-025-00516-2
- Faizi, M., Qiang, Y., Wei, Y., Qiao, Y., Zhao, J., Aftab, R., & Urrehman, Z. (2025). Deep learning-based lung cancer classification of CT images. *BMC Cancer*, 25, 1056. doi:10.1186/s12885-025-14320-8
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2019). Synthetic data augmentation using GAN for improved liver lesion classification. *IEEE Transactions on Medical Imaging*, 38(3), 677-685. doi:10.1109/TMI.2018.2865516
- Gai, L., Xing, M., Chen, W., et al. (2024). Comparing CNN-based and transformer-based models for identifying lung cancer: Which is more effective? *Multimedia Tools and Applications*, 83, 59253-59269. <https://doi.org/10.1007/s11042-023-17644-4>
- Goswami, A., Chandra, K., & Kar, S. (2026). A Comprehensive Multi-Model Framework for Lung Cancer Image Classification Integrating Advanced Deep Learning and Explainable AI. 2026 9th International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), (pp. 1-6).
- Hammad, M., El-Affendi, M., El-Iatifi, A., Ateya, A., Ali, G., & P?awiak, P. (2025). Explainable AI for lung cancer detection via a custom CNN on CT images. *Scientific Reports*.
- Haque, F., Hasan, M. A., Siddique, M. A. I., Roy, T., Shaha, T. K., Islam, Y., ... & Chowdhury, M. E. (2025). An End-to-End Concatenated CNN Attention Model for the Classification of Lung Cancer With XAI Techniques. *IEEE Access*.
- Hatamizadeh, A., Tang, Y., Nath, V., & Yang, D. (2022). UNETR: Transformers for 3D medical image segmentation., (pp. 574-584). doi:10.1109/WACV51458.2022.00063
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 770-778).
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. (2017). Densely Connected Convolutional Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 4700-4708).
- Javed, R., Abbas, T., Khan, A., Daud, A., Bukhari, A., & Alharbey, R. (2024). Deep learning for lungs cancer detection: A review. *Artificial Intelligence Review*, 57, 197. doi:10.1007/s10462-024-10807-1
- Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.
- Lung Cancer. (2024). Retrieved from <https://www.who.int/news-room/fact-sheets/detail/lung-cancer>

- Ma, K., Zheng, M., Chen, W., Qi, Y., & Rong, H. (2025). Research progress in computer-aided diagnosis systems for lung cancer. *npj Digital Medicine*, 8, 722. doi:10.1038/s41746-025-02101-3
- Mannepalli, D., Tak, T. K., Krishnan, S. B., & Sreenivas, V. (2025). GSC-DVIT: A vision transformer based deep learning model for lung cancer classification in CT images. *Biomedical Signal Processing and Control*, 103, 107371. <https://doi.org/10.1016/j.bspc.2024.107371>
- Nithya, R., & Vidhyapathi, C. (2025). Transfer learning based deep architecture for lung cancer classification using CT image with pattern and entropy based feature set. *Scientific Reports*, 15, 28283. doi:10.1038/s41598-025-13755-0
- Oktay, O., Schlemper, J., Le Folgoc, L., & Lee, M. (2018). Attention U-Net: Learning where to look for the pancreas. Retrieved from <https://arxiv.org/abs/1804.03999>
- Oncu, E., & Ciftci, F. (2025). Multimodal AI framework for lung cancer diagnosis: Integrating CNN and ANN models for imaging and clinical data analysis. *Computers in Biology and Medicine*, 193, 110488.
- Revathi, T., & Leni, A. E. S. (2025). Hybrid transformer-based framework for multi-type lung cancer detection using attention-guided feature fusion. *Procedia Computer Science*, 270, 4533–4542. <https://doi.org/10.1016/j.procs.2025.09.579>
- Selvaraju, R., Cogswell, M., & Das, A. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *IEEE International Conference on Computer Vision*, (pp. 618-626).
- Setio, A., Traverso, A., & De Bel, T. (2017). Validation, Comparison, and Combination of Algorithms for Automatic Detection of Pulmonary Nodules in Computed Tomography Images: The LUNA16 Challenge. *Medical Image Analysis*, 42, 1-13.
- Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2017). Multi-scale convolutional neural networks for lung nodule classification. (pp. 588-599). Springer. doi:10.1007/978-3-319-59050-9_47
- Siegel, R., Giaquinto, A., & Jemal, A. (2024). Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 74(1), 12-49.
- Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. 97, pp. 6105-6114. PMLR.
- Wang, M., Yang, Z., & Zhao, R. (2025). Advancing lung cancer diagnosis: Combining 3D auto-encoders and attention mechanisms for CT scan analysis. *Journal of X-Ray Science and Technology*, 33(2), 376–392. <https://doi.org/10.1177/08953996241313120>
- Woo, S., Park, J., Lee, J., & Kweon, I. (2018). CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision*.
- Zhou, L., Jain, A., Dubey, A., Singh, S., & Gupta, N. (2025). FPA-based weighted average ensemble of deep learning models for classification of lung cancer using CT scan images. *Scientific Reports*, 15, 19369. doi:10.1038/s41598-025-02015-w
- Zhou, Z., Siddiquee, M., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. (pp. 3-11). Springer. doi:10.1007/978-3-030-00889-5_1