Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HYBRID CNN TRANSFORMER MODELS FOR LUNG DISEASE DETECTION FROM CT SCANS

ARUNA VIPPARLA[1], DR. LAKSHMI NAGA JAYAPRADA GAVARRAJU[2], DR. B. LEELAVATHY[3], DR. JOHNWESILY CHAPPIDI[4], CHALLAPALLI SUJANA[5], KUMAR DEVAPOGU[6]*, AKKALA YUGANDHARA REDDY[7], K. PRAVEEN KUMAR[8]

[1]Assistant Professor, Department of CSE, Dr RVR NRI Institute of Technology Deemed to be University, Vijayawada, Andhra Pradesh, India
[2]Sr. Assistant Professor, Department of CSE (CyS, DS) and AI&DS, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology, Hyderabad, Telangana, India
[3]Assistant Professor, Department of Information Technology, Vasavi College of Engineering, Hyderabad, Telangana, India
[4]Associate Professor, Department of CSE (AI&ML), Lakireddy Bali Reddy College of Engineering (A), Mylavaram, Andhra Pradesh, India
[5]Assistant Professor, Department of CSE, Aditya University, Surampalem, Andhra Pradesh, India - 533437
*[6]Assistant Professor, Department of CSE, Vignan's Foundation for Science, Technology & Research, Guntur, Andhra Pradesh, India
[7]Assistant Professor, Department of CSE, Vignan's Nirula Institute of Technology and Science for Women, Guntur, Andhra Pradesh, India
[8]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

E-mail: [1]aruna.vipparla5@gmail.com, [2]jayaprada_g@vnrvjiet.in, [3]leelapallava@staff.vce.ac.in, [4]wesily013@lbrce.ac.in, [5]sujana.challapalli@gmail.com, *[6]dk_cse@vignan.ac.in, [7]akkalayugandhara20@gmail.com, [8]pravi.kanapala@gmail.com

ABSTRACT

Timely and accurate diagnosis is significant for enhancing patient survival and reducing the strain that health care facilities would otherwise face in providing diagnostic services. Computed Tomography (CT) imaging is a popular imaging method that allows visualization of lung structure in detail, revealing abnormalities, including lung cancer, pneumonia, and Coronavirus (COVID-19) infection. Nevertheless, manual interpretation of CT scans is labor-intensive and susceptible to discrepancies among radiologists. The work presents an effective automated system for lung disease detection using a hybrid CNN-Transformer architecture that utilizes local spatial attributes and extensive contextual comprehension. The convolutional neural networks (CNNs) in this paper formulate a model that extracts multi-level features, after which a Transformer encoder is used to model long-range dependencies across lung regions. Moreover, a new multi-scale attention fusion framework is proposed to combine the spatial and contextual representations to enhance the feature learning. The experiment was conducted on a multi-class lung CT dataset that includes normal cases, lung cancer, pneumonia, and COVID-19. The proposed model achieved a classification accuracy of 96.3%, a sensitivity of 95.8%, a precision of 95.9%, and an ROC-AUC of 0.97, surpassing other conventional CNNs, including VGG16, ResNet50, and DenseNet121, as well as single Vision Transformer models. The results show that the hybrid CNN-Transformer architecture outperforms both CNN and Transformer architectures in feature encoding and lung disease diagnosis. The proposed framework constitutes a reliable computer-aided diagnostic system that can assist radiologists in the initial identification of pulmonary diseases and has great potential for use in AI-assisted clinical decision support systems.

Keywords: *Lung Disease Detection, Computed Tomography (CT), Hybrid CNN–Transformer, Multi-scale Attention Fusion, Deep Learning, Medical Image Classification*

1. INTRODUCTION

The lung diseases remain a highly important health concern around the globe, and they are also among the leading causes of death on our planet [1]. Other diseases, such as lung cancer, pneumonia, tuberculosis, and viral respiratory infections, have been reported to contribute

immensely to morbidity and medical costs in the world [2]. Of all these, lung cancer alone leads to high numbers of cancer-related deaths due to the delay in its diagnosis as well as the unavailability of the means of prompt detection [3]. Early detection of pulmonary abnormalities is therefore relevant for achieving a high survival rate and an effective response to the abnormalities. Medical imaging procedures such as computed tomography (CT) scanning provide high-quality images of internal lung structures and have become a diagnostic tool for detecting abnormalities in lung tissue [4].

Computed Tomography provides high-resolution sectional images that help clinicians identify minor pathological changes, e.g., nodules, infiltrates, and lesions. CT scans are more accurate in spatial information and may be better able to diagnose lung diseases at an earlier stage, which would otherwise be hard to detect with conventional chest X-rays [5]. However, the rising costs of CT imaging in modern medical records pose enormous challenges for radiologists [6]. Interpretation of CT scans is tedious and likely to involve interobserver variation, leading to inconsistent diagnoses and errors [7]. Interpretation of CT scans is a subjective, labour-intensive, and time-consuming process, with the potential to affect diagnostic consistency and accuracy when performed manually by different interpreters [8].

However, as the last few years demonstrated, artificial intelligence (AI) and deep learning methods can be used to address them and analyse medical images automatically. Various Computer vision problems, such as image classification, segmentation, and object detection, have been implemented more effectively with deep learning models [9]. Convolutional Neural Networks (CNNs) have begun to dominate the field of medical image analysis, as they can learn multi-level feature representations directly from raw images without feature extraction [10], [11]. Lung disease detection using CNN-based models has been widely applied for CT images. Such models can effectively learn local spatial receptive fields, including lung nodules, lesions, and tissue abnormalities, using convolutional filters. A number of studies have demonstrated that CNN architectures, such as AlexNet, VGGNet, ResNet, and DenseNet, can accurately identify pulmonary diseases [12], [13]. On the one hand, residual networks (ResNets), such as adding residual links to overcome vanishing gradients and using them more often in deeper networks for enhancing feature extraction [14], were introduced. Likewise, DenseNet models improve feature propagation by connecting every layer to all other layers in a feed-forward manner, consequently enhancing gradient propagation and model compression [15].

Although they were successful, CNN-based models lack certain characteristics for handling complex tasks in medical images. Among the primary drawbacks of CNNs is their inability to capture long-range dependencies in images. This may also make CNNs less able to learn global contextual connections between remote image features, since the discussion of convolutional operations is primarily in terms of local receptive fields. Disease patterns on lung CT scans are often scattered across multiple lung regions; therefore, there is a dire need to ensure that the global context is captured to achieve proper diagnosis [16].

To address this shortcoming, an attention mechanism has been added to deep learning models to enhance feature representation by highlighting the most important parts of an image. Selective attentive CNN models have demonstrated better performance on medical image-related tasks by extracting significant and relevant spatial differences in disease patterns [17]. Nevertheless, attention models can enhance the weighting of local features but also rely on convolutional architectures that, in turn, limit the modeling of global correlations.

Transformer architecture tools have recently become direct competitors to long-range relation modeling for sequential and image data. Transformers were originally formulated for natural language processing using self-attention methods that model relationships between items in the input data irrespective of their physical distance. This feature was introduced with the advent of ViT models (ViT) and is now used in image analysis applications, which divide images into patches and learn global contextual connections via multi-head self-attention layers [18]. Transformers have achieved state-of-the-art results on a wide range of computer vision benchmarks and have shown strong performance in medical imaging classification [19].

Transformer-driven architectures, however, also face certain challenges in medical imaging. Transformers generally require a large training dataset and considerable computational power to perform well. Medical imaging datasets are more likely to be small due to confidentiality issues, the cost of data annotation, and scarce access to adequate data. Consequently, the standalone Transformer models can fail to generalize well when trained using rather small datasets [20].

To address the shortcomings of CNN models, recent studies have investigated densely coupled CNN-Transformer architectures that use CNNs for feature extraction plus Transformers for contextual representation. The successful learning of low- and mid-level spatial attributes in these combined architectures is achieved by CNN layers, and overall contextual association and long-range dependencies between image sections are achieved by Transformer encoders. The blended application of hybrid models enables them to capitalize on the advantages of both architectures while lessening their shortcomings [21].

Despite significant advances in medical image analysis using deep learning, multi-class lung disease detection from computed tomography (CT) scans remains a challenging problem due to disease patterns, inter-class similarity, and the need to capture both local and global features in medical images. Traditional CNN models work well for local spatial patterns, such as texture changes, lesions, or nodules, but they fail to capture context-dependent information spanning large organs, such as the lung or parts thereof, that is required to capture the complex nature of diseases. Transformer models, on the other hand, have demonstrated excellent contextual awareness globally via self-attention mechanisms but often require massive amounts of annotated data and high-performance demands, making them difficult to adopt in medical imaging applications.

The CNN-T has achieved remarkable results across a range of computer vision tasks, including object detection, medical image segmentation, and disease classification. In medical imaging, such models have been shown to enhance doctors' diagnostic correctness by enabling them to combine the details of spatial representations with general information across the entire image. However, the available hybrid models primarily focus on general image analysis and therefore do not require lung CT image-specific features. Moreover, scale-based feature combination strategies, which can effectively integrate spatial and contextual representations to detect lung disease, have received limited research attention [22].

Recently, some hybrid CNN–Transformer approaches have been proposed. However, most current methods address general medical image analysis but do not specifically target multi-class lung disease classification from CT scans. Few studies integrate multi-resolution spatial and contextual information to detect subtle variations and ensure reliable diagnosis. Prior work also notes the difficulty of distinguishing between diseases with similar radiological signs, such as Pneumonia and COVID-19.

An intelligent framework with efficient data capacity, the ability to learn local features with CNNs (Convolutional Neural Networks), global context with Transformers (deep learning models using self-attention for contextual relationships), and an effective feature fusion mechanism (combining features learned by different models to produce improved representations) to enhance disease representation is strongly desired. To address these limitations, a hybrid CNN–Transformer model with Multi-Scale Attention Fusion is proposed to improve feature map learning, achieving higher classification accuracy and creating a robust computer-aided diagnostic system for early and accurate diagnosis of lung diseases from CT images.

This research was driven by the issues raised above, and therefore, the researchers introduced a Hybrid CNN-Transformer architecture to detect lung diseases upon receipt of the CT scan. This system consists of a backbone deep CNN to extract multi-level features and a Transformer encoder to encode global contextual connections between regions, e.g., the lung. Also, a multi-scale attention fusion method is proposed to improve feature representation, where features are encoded with spatial and contextual data acquired at multiple scales within the network.

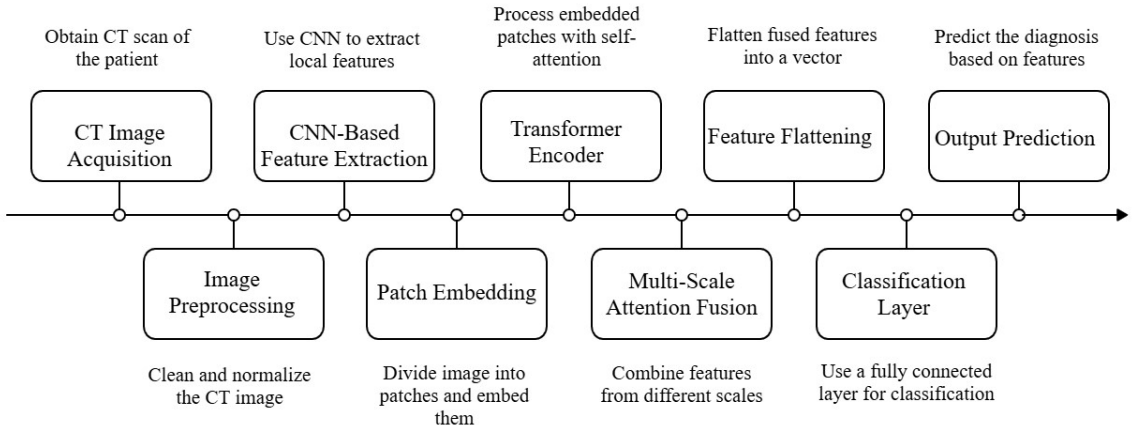

Figure 1. *Hybrid CNN-Transformer Architecture Workflow*

The workflow involves a hybrid CNN-Transformer model that classifies lung diseases, as depicted in Figure 1. The model involves CT image acquisition, pre-processing, and feature extraction using CNNs that detect local patterns. These features are subsequently encoded into patches and transduced by a Transformer encoder to acquire contextual knowledge. A multi-scale attention fusion unit improves the features, which are then flattened and sent to the classification layer to produce the final diagnosis.

The main objective of this study was to develop an advanced hybrid deep learning architecture based on Convolutional Neural Networks (CNNs) and Transformer networks to efficiently process lung CT images. The KP model was intended to improve the accuracy of lung disease detection through integrating the high spatial locality of CNNs with the worldwide contextual representation of Transformer-based self-attention. Moreover, the multi-scale attention fusion strategy was included in the suggested strategy, meant to enhance feature encoding by providing information from spatial and contextual domains across layers of our network. Lastly, the research evaluated the performance of the proposed hybrid architecture relative to conventional CNN- and Transformer-based models using commonly used evaluation measures in medical imaging, including accuracy, precision, recall, F1, and ROC-AUC.

The proposed framework intends to create an effective computer-aided diagnostic system that helps radiologists examine CT scans to detect lung diseases more accurately and reliably. The model will also integrate the strengths of CNN and Transformer models for enhancing automatic detection of lung-related diseases and further advance AI-based healthcare technology [23].

The rest of this paper is organized as follows. Section 2 evaluates existing research on deep learning methods for lung disease detection and on CNN-Transformer hybrid architectures in medical imaging. Section 3 presents the suggested methodology, including descriptions of the data used, the data preprocessing methods, the hybrid CNN-Transformer model, the mathematical model, and the proposed algorithm. The experimental results and performance evaluation presented in Section 4 are based on traditional assessment measures and contrast them with the existing state-of-the-art models. Lastly, Section 5 draws conclusions from the paper, presenting the main findings, the study's limitations, and proposals for future research.

2. RELATED WORK

2.1 CNN-Based Approaches for Lung Disease Detection

Convolutional Neural Networks (CNNs) based on deep learning have been widely used for the automatic analysis of medical images [24]. There is also significant success with CNN-based architectures for detecting lung diseases in CT scans [25]. This is because of their ability to acquire hierarchical spatial properties from raw images. The first research used deep CNN models to identify pulmonary nodules and lung cancer in the CTs [26]. These models showed their ability to learn discriminative controls for lung abnormalities. Consequently, they are more effective for classification than conventional machine learning methods [27].

Based on those advancements, other trending CNNs, such as VGGNet, Inception, and ResNet, have also been used to classify lung CT images [28]. These networks can be used to describe spatial features, such as edges, textures, and lesion patterns, using lung images [29]. For example, lung

nodules have been identified using a deep residual learning model that leverages deep network structures via shortcut connections to avoid the vanishing gradient problem [30]. These research studies reported higher detection performance and sensitivity to lung abnormalities [31].

However, despite their success, CNN-based models are likely to be unsuitable for retrieving long-range contextual relations within an image [32]. Since convolutional functions operate on local receptive fields, the models might miss global functions that could be significant for identifying detailed patterns of disease across large regions of the lung.

2.2 Multi-Scale CNN Architectures

Multiscale CNNs aim to address the inefficiency of conventional CNNs in producing a variety of features at different spatial scales. The concept of multiscale learning enables the network to acquire finer-scale localisation patterns and learn coarse-scale background information from CT scans [33]. Based on this Development, various researchers have recommended multi-branch CNNs for CT images at different scales. These buildings integrate feature maps from two or more convolutional layers, enabling the model to detect both small nodules and larger lung defects. The multiscale nature of these structures is more sensitive for the early detection of lung diseases, when malfunctions might be minor in lung tissues [34]. In addition to multi-branch approaches, pyramid-based CNN architectures have been designed to better represent the features to combine multi-level spatial information. This approach improved the detection of small pulmonary nodules and subtle disease patterns that are difficult to detect with single-scale CNN models [35].

2.3 Attention-Based Deep Learning Models

Attention mechanisms have become an effective technique for enhancing the performance of deep learning models for medical imaging tasks [36]. (Automated classification of chest X-rays: a deep learning approach with attention mechanisms, 2025) Attention modules encourage networks to focus on the most relevant regions of an image, thereby enhancing feature representation and making it more interpretable [37]. Spatial and channel attention mechanisms have been incorporated into CNN frameworks for detecting lung diseases and highlighting disease areas in CT scans [38]. These attention-based applications indicate to the network the features pertinent to lung disease and pathologies (such as nodules, infiltrates, or ground-glass opacities) [39]. It has been demonstrated that attention, combined with CNN architectures, can enhance classification accuracy while reducing the false-positive rate. An attention-guided CNN model also makes deep learning systems much easier to understand by producing visual attention maps that show which parts of the input influence the model's predictions [40].

2.4 Transformer-Based Models in Medical Imaging

Over the past few years, transformer architectures have become widely popular in computer vision hardware because they can capture global contextual relationships via their self-attention mechanisms [41]. The Vision Transformer (ViT) model represents images as patches, and multi-head self-attention is used to model relationships between different parts of the image. The use of Transformer models in medical image classification has been covered in several studies, including those that consider lung CT images [42]. Architecture based on transformers has demonstrated strong performance in learning global context, a challenging task for CNNs [43]. These models may be useful for discovering correlations between remote regions of CT scans, which is significant for identifying distributed patterns of disease throughout the lung. Transformer models, however, require both large datasets and substantial computing power to be trained successfully [44]. Challenges in obtaining medical imaging databases are also a general issue, as annotating medical images is expensive, and privacy concerns may compromise the performance of Transformer models in single-user settings on clinical datasets [45].

2.5 Hybrid CNN–Transformer Architectures

To limit the use of both CNNs and Transformers, researchers have, in recent years, developed hybrid models that incorporate them. Hybrid CNN-Transformer systems will also leverage the local extraction features of CNNs, the long-distance relationships of a transformer network, and its global overview.

In this architecture, CNN modules are commonly used as backbone networks to extract hierarchical features from medical images. Transformer layers are then applied to them, effectively modeling interactions between various regions of the image, yielding self-attention results. It has been demonstrated that hybrid architectures outperform single-CNN or single-Transformer architectures across a variety of computer vision tasks [46].

Medical Imaging: Hybrid CNNs and Transformers have been used for tasks such as

tumor detection, organ segmentation, and disease classification. These models can synthesize spatial and contextual representations and provide enhanced diagnostics and greater robustness [47].

2.6 Multi-Scale Feature Fusion and 3D CT Analysis

The other important trend in research is multi-scale feature fusion in deep learning models. Multi-scale fusion techniques leverage features from multiple network depths and integrate complementary features at varying levels of blurriness. Several studies have demonstrated that multi-scale feature fusion incorporating low-level texture and high-level semantic features improves the detection of subtle lung abnormalities [48]. This process is further improved by attention-based fusion mechanisms that assign adaptive weights to features at various scales [49]. Also, researchers have developed 3D deep learning architectures that not only analyze single slices but also volume CT data. A 3D CNN model can leverage the spatial relationships among cuts from multiple CTs, thereby providing more contextual information about lung diseases [50]. However, such models need much more CPU time and memory, which is a limitation for real-time clinical implementations [51]. Consequently, hybrid architectures integrating effective 2D CNN-based feature extractors with Transformer-based contextual modelling have been increasingly regarded as viable solutions for lung disease detection [52].

3. METHODOLOGY

The proposed framework presents a Hybrid CNN-Transformer architecture with Multi-Scale Attention Fusion for automatically detecting lung disease in CT scans. The model uses convolutional feature extraction along with transformer-based contextual learning to identify both local spatial patterns and long-range relationships in lung CT images.

3.1 Dataset Description

Publicly available lung CT data were utilized to test the proposed framework. The datasets consist of CT scans of various lung diseases, including healthy lungs, pneumonia, lung cancer, and COVID-19.

The search engines used were the MosMedData, the COVID-CT Dataset, and the Lung Image Database Consortium (LIDC-IDRI), all of which offer free medical databases. High-resolution CT scans from various hospitals and imaging systems are collected in the datasets.

The data were collected from three widely recognized repositories: MosMedData [53], COVID-CT Dataset [54], and LIDC-IDRI [55]. These datasets contain high-resolution CT scans from multiple hospitals and imaging systems. This ensures variability and robustness for model training and evaluation.

Table 1. *Dataset Parameters*

Parameter	Value
Total CT Images	12,000
Image Modality	Computed Tomography
Image Format	PNG / DICOM
Image Resolution	512 × 512
Resized Input Size	224 × 224
Number of Classes	4
Training Samples	8,400
Validation Samples	1,800
Testing Samples	1,800

According to Table 1, 12,000 Computed Tomography (CT) images are included in the dataset of this study, as they provide sufficient variability and classification capabilities. The images also come in PNG and DICOM formats, making them easily preprocessed and conforming to medical imaging standards. All images initially have a pixel resolution of 512 x 512, which maintains fine-grained anatomical details vital to proper diagnosis. To ensure efficient computation and meet the input requirements of deep learning models, all images are resized to 224 x 224 before training.

The dataset is divided into four classes representing various disease conditions, making it suitable for multi-class classification. In order to have an appropriate training and evaluation, the dataset will be split into three subsets: 8,40In order to have an appropriate training and evaluation, the dataset will be split into three subsets: 8,400 images will be used in the training, 1,800 images in the validation, and the remaining 1,800 in the testing. Evaluation of performance will be reliable. In general, the dataset is well-organized and large enough to design and test highly advanced deep learning models in medical image analysis.

To evaluate the model's performance without bias, the dataset was split into 70% for training, 15% for validation, and 15% for testing.

Table 2. *Class Distribution*

Class	Number of Images
Normal	3,000
Lung Cancer	3,000
Pneumonia	3,000
COVID-19	3,000

The dataset is evenly distributed across four classes (Table 2), with 3000 images per class, to ensure balanced representation for multi-class classification. The normal class includes CTs with

no pathological abnormalities and serves as a control. The Lung Cancer class includes images of malignant nodules or patterns of tumor growth, which are useful for early cancer diagnosis. The Pneumonia category includes images of lung opacities and infection-related inflammation, as well as ground-glass opacities and bilateral lung involvement on CT images, which are covered by the other class, which in turn is the class of patients with coronavirus (Cov-19).

This method of 3,000 images per class could prevent the model from training on the class imbalance issue, which is significant because it supports unbiased model training and fairness in evaluating the emerging model's performance. It also helps the model acquire unique features in each condition, thereby improving classification and generalization. All in all, the balanced dataset structure makes the proposed deep learning model reliable and robust.

3.2 Image Preprocessing

Medical CT images are often noisy, contain inappropriate background structures and exhibit intensity variations that can affect the model's performance. Thus, several preprocessing operations were used.

Preprocessing Steps

i. Intensity Normalization

Pixel intensities were normalized using min–max normalization:

$$I_{norm}=\frac{I-I_{min}}{I_{max}-I_{min}} \quad (1)$$

Where: Irepresents the original pixel intensity.

ii. Noise Reduction

A Gaussian filtering operation was applied to remove high-frequency noise.

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

iii. Lung Region Segmentation

Threshold-based segmentation was used to isolate lung regions from surrounding tissues.

iv. Image Resizing

All CT images were resized to:

$$224\times224 \quad (3)$$

to match the CNN input size.

v. Data Augmentation

To increase dataset diversity and prevent overfitting, augmentation techniques were applied.

Table 3. *Image Data Augmentation Techniques and Parameter Settings*

Augmentation Technique	Parameter
Rotation	±20°
Horizontal Flip	Enabled
Zoom	0.2
Translation	10%

The data augmentation techniques were tested in Table 3 to ensure the training data set is diverse and robust. The level of rotation is varied in the range of +20 to -20; after which the model becomes orientation invariant. Horizontal 16x flipping has been activated to simulate anatomical variability and diversify the mirror and sample sets. Zoom: This variable affects learning; therefore, you can adjust your model to be sensitive to scale changes by setting the zoom factor to 0.2. Moreover, a maximum of 10 per cent translation is done to create horizontal and vertical lines to fit the location differences of key parts. These data augmentation techniques reduce overfitting, improve overall generalization, and sequentially enhance the performance of deep learning models.

3.3 Proposed Hybrid CNN–Transformer Architecture

The architecture proposed in the study integrates CNN layers and Transformer encoders to capture both local and global spatial and contextual associations.

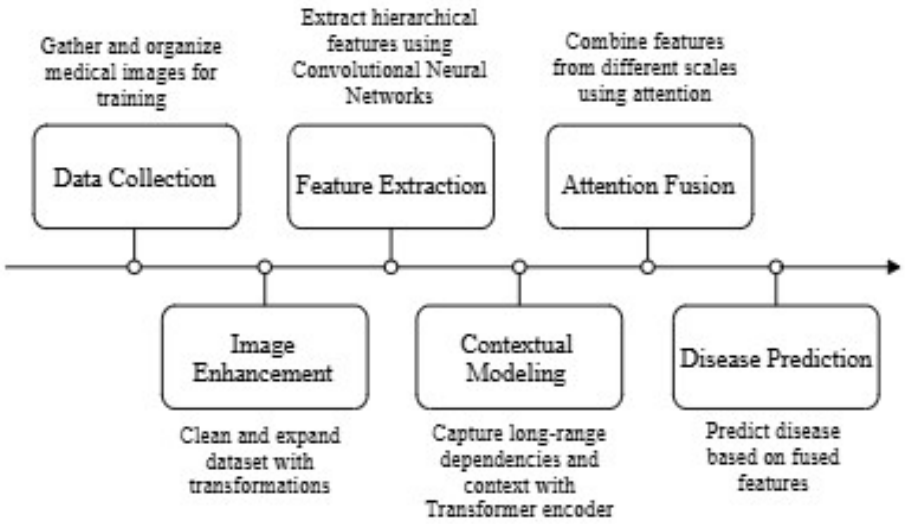

Figure 2. *Proposed System Workflow for Disease Prediction*

The proposed system is designed with a clear chain of actions that constitute the workflow, enabling reproducibility. Figure 2 shows that the first phase is the preparation and acquisition of the dataset, including obtaining appropriate medical images from reliable sources, annotating them, normalizing them, and splitting them into training, validation, and test sets. Image preprocessing and enhancement mechanisms such as resizing, noise removal, contrast enhancement, rotation, flipping, and scaling are then used to improve data quality and increase data diversity. The second step involves a hierarchical feature extraction mechanism that uses a Convolutional Neural Network (CNN) to encode low-level to high-level spatial features of the input images. These features are later sent to a Transformer encoder, which captures long-range dependencies and contextual relationships among the feature maps. In a further move to improve representation, it is proposed that a multi-scale attention fusion module, which integrates features across scales, be used to enable the model to attend to the most informative areas. Lastly, the processed feature image is fed into a fully connected classification layer, which produces the final disease prediction. All these stages are well-defined and detailed to ensure the proposed system is reproducible and performs well.

The proposed architecture comprises five major modules that perform specific tasks throughout the learning process. At first, the CNN feature extraction module can capture hierarchical spatial features from the input images, namely, identifying low-level features, such as edges and textures, and high-level semantic features. These extracted features are then passed to the patch embedding module, where the feature map is split into smaller patches and converted into a series of feature mappings, ready to be fed into a transformer. The transformer encoder models long-range dependencies and contextual relationships among patches, enabling a better understanding of the image's global features. To further enrich feature representation, a multi-scale attention fusion module is employed to combine information across multiple scales, enabling the model to attend to the most relevant regions at different resolutions. Finally, the classification layer uses the fused features and outputs the predicted class label, completing the disease classification task.

3.3.1 CNN Feature Extraction

A deep CNN backbone is used to extract hierarchical spatial features.

The convolution operation is defined as:

$$F(i,j)=(X*K)(i,j)=\sum_m \sum_n X(i-m,j-n)K(m,n) \quad (4)$$

where:

- X= input image
- K= convolution kernel
- $F(i,j)$= feature map

The CNN also includes a series of convolutional and pooling layers that can extract meaningful spatial features from medical images. The layers can identify meaningful features of images, e.g., lung nodules, variations in tissue texture and correct lesion boundaries, all of which are essential to successful disease detection and classification.

Table 4. *CNN Layer Configuration*

Layer	Filters	Kernel Size
Conv1	32	3×3
Conv2	64	3×3
Conv3	128	3×3
MaxPooling	2×2	—

The CNN layers in Table 4 are used for feature extraction. It uses three convolutional layers with filters of increasing sizes (32, 64, 128) and a 3x3 kernel to extract spatial characteristics, followed by a 2x2 max-pooling layer to reduce dimensions and select essential features.

3.3.2 Patch Embedding

CNN feature maps are split into patches or patch tokens, which are then fed into the Transformer. The patches are flattened and represented as vectors.

$$z_0=[x_1E;x_2E;\ldots;x_NE] \quad (5)$$

where:

- x_i= image patch
- E= embedding matrix
- z_0= patch embedding sequence

3.3.3 Transformer Encoder

Transformer encoders learn long-range interaction through multi-head self-attention.

Self-Attention Mechanism

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

where

- Q= Query matrix
- K= Key matrix
- V= Value matrix
- d_k= scaling factor

Multi-Head Attention

$$\text{MHA}(Q,K,V)=\text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \quad (7)$$

Through this mechanism, the network can acquire global contextual information about CT images.

3.4 Multi-Scale Attention Fusion Module (Novel Contribution)

A Multi-Scale Attention Fusion (MSAF) module is added to enhance feature representation. The fusion process incorporates the features that have been extracted from:

- CNN layers (local features)
- Transformer layers (global context)

Feature fusion is defined as:

$$F_{fusion}=\alpha F_{cnn}+\beta F_{transformer} \quad (8)$$

where:

- F_{cnn}= spatial feature map
- $F_{transformer}$= contextual feature map
- α, β= adaptive attention weights

The weights are dynamically updated during training. This module enhances the detection of subtle lung anomalies.

3.5 Classification Layer

The resulting fused feature vector is sent to fully connected layers for classification.

$$y=\text{softmax}(Wx+b) \quad (9)$$

where:

- W= weight matrix
- b= bias
- y= predicted probability

The output corresponds to one of the four classes:

- Normal
- Lung Cancer
- Pneumonia
- COVID-19

3.6 Proposed Algorithm

Algorithm: Hybrid CNN–Transformer Lung Disease Detection
Input: CT scan image dataset
Output: Predicted lung disease class

a) Load lung CT dataset
b) Apply preprocessing (normalization, filtering, segmentation)
c) Resize images to 224 × 224
d) Apply data augmentation techniques
e) Feed images into CNN feature extractor
f) Generate hierarchical feature maps
g) Convert feature maps into patch embeddings
h) Pass embeddings to Transformer encoder
i) Compute multi-head self-attention
j) Extract contextual features
k) Apply Multi-Scale Attention Fusion module
l) Flatten fused feature representation
m) Pass features to fully connected classifier
n) Apply softmax activation to obtain class probabilities
o) Output predicted lung disease class

3.7 Novel Contributions of the Proposed Method

The proposed research approach will include the inclusion of CT image analysis to boost the efficiency of lung CT image analysis. It is trained on a hybrid CNN-Transformer architecture that excels at integrating spatial features and context to improve diagnosis. It also contains a multi-scale attention fusion module that consolidates features across different resolutions, helping the model refine not only fine-grained but also global information. Moreover, an adaptive attention-weighting framework is introduced that targets disease features to enhance classification across classes. In addition, the feature-embedding tactic is well utilised to advance contextual modelling and augment the depiction of intricate trends in the data.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

These innovations are combined to form a superior rubric for classifying diseases.

4. RESULTS AND DISCUSSION

This paragraph presents the experimental analysis of the proposed Hybrid CNN-TS Transformer with Multi-Scale Attention Fusion (HCT-MSAF) model for detecting lung diseases using CT images. The model's performance is evaluated using a variety of metrics and compared with state-of-the-art deep learning methods. Experiments were conducted using the dataset described in the methodology section. All experiments were split into 70, 15 and 15 percent for training, validation and testing, respectively.

4.1 Experimental Setup

The PyTorch deep learning manual was used to implement the proposed model, which was trained on a workstation with a provided GPU. Training was done by using mini-batch stochastic gradient descent with adaptive optimization.

Table 5. *Training Parameters*

Parameter	Value
Framework	PyTorch
Batch Size	32
Optimizer	Adam
Learning Rate	0.0001
Number of Epochs	50
Loss Function	Cross-Entropy Loss
Input Image Size	224 × 224

Table 5 gives the training setup of the proposed model. It is trained with PyTorch, a batch size of 32, and the Adam optimizer with a learning rate of 0.0001. Cross-Entropy Loss is used to train the model using 50 epochs, and the input images are resized to 224 (square) to maintain consistency. Early stopping and dropout regularization were also used to prevent overfitting.

4.2 Evaluation Metrics

A complete performance evaluation was required, so the proposed model was assessed using the standard classification metrics commonly used in medical imaging research.

Accuracy

Accuracy measures the overall classification performance:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Precision

Precision indicates the proportion of correctly predicted positive cases.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (11)$$

Recall (Sensitivity)

Recall measures the ability of the model to correctly detect disease cases.

$$\text{Recall}=\frac{TP}{TP+FN} \quad (12)$$

F1-Score

The F1-score represents the harmonic mean of precision and recall.

$$F1=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (13)$$

Specificity

Specificity measures the ability to correctly identify normal cases.

$$\text{Specificity}=\frac{TN}{TN+FP} \quad (14)$$

ROC-AUC

The Receiver Operating Characteristic (ROC) curve measures the model's classification performance at varying decision thresholds.

4.3 Training Performance Analysis

During training, the model experienced consistent convergence, with the loss decreasing and classification accuracy increasing.

Table 6. *Training and Validation Accuracy*

Epoch	Training Accuracy	Validation Accuracy
10	89.2%	88.4%
20	92.6%	91.9%
30	94.5%	93.8%
40	95.8%	95.2%
50	**96.9%**	**96.3%**

Table 6 shows that the training and validation accuracies improve progressively with increasing epochs, reaching 96.9% and 96.3% after 50 epochs. The small separation between them implies it is possible to generalize well and minimize overfitting.

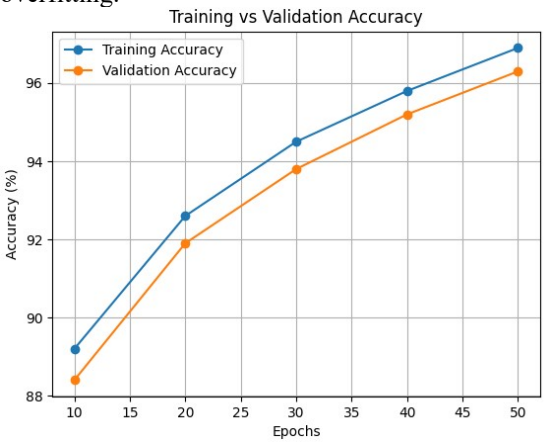

Figure 3. *Training and Validation Accuracy Curves Demonstrating Model Convergence and Minimal Overfitting*

Training and validation accuracies are compared at various epochs (10-50), as shown in Figure 3. The training and validation accuracies have continued to increase with each epoch, indicating the model is learning. The training

accuracy is 89.2 per cent, increasing to approximately 96.9 per cent, and the validation accuracy is 88.4 per cent, increasing to 96.3 per cent. The difference between the training and validation accuracies is small during training, indicating that the model is generalizable and not overfitting. Stability in learning and high model performance indicate steady, parallel improvement in the two curves. Overall, the more epochs, the higher the accuracy, and the model performs quite well, demonstrating a high degree of generalization. Training and validation curves indicate that the proposed architecture can significantly reduce overfitting.

4.4 Classification Results

The ultimate analysis of the test data showed a high classification rate across all categories of lung diseases.

Table 7. *Overall Performance Metrics*

Metric	Proposed Model
Accuracy	**96.3%**
Precision	**95.9%**
Recall (Sensitivity)	**95.8%**
F1-Score	**95.8%**
Specificity	**96.9%**
ROC-AUC	**0.97**

Table 7 indicates that the proposed model performs well, achieving high accuracy, precision, recall, and F1-score (96.3% and 95.8%, respectively). Specificity (96.9%) and ROC-AUC (0.97) are good and indicate very reliable and discriminating performance. The high ROC-AUC value indicates strong discrimination by the model.

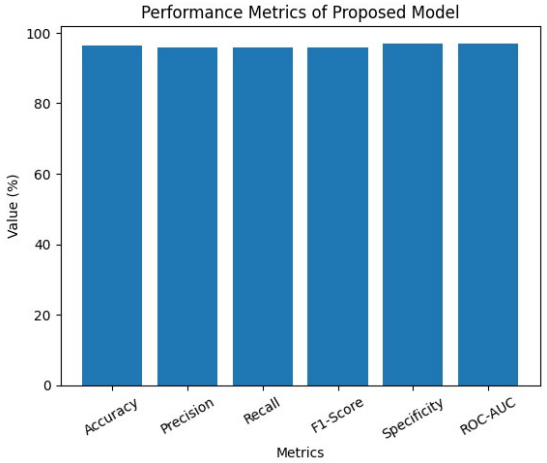

Figure 4. *Performance Evaluation Metrics of the Proposed Model*

As Figure 4 indicates, the proposed model has high performance across all measures, with an accuracy of 96.3 percent and strong precision, recall, and F1-score (approximately 95.8 percent). The specificity (96.9%) and ROC AUC (0.97) are high, indicating high reliability and classification performance.

4.5 Confusion Matrix Analysis

The confusion matrix illustrates the classification results for each disease category.

Table 8. *Confusion Matrix of the Proposed Model for Multi-Class Lung Disease Classification*

Actual / Predicted	Normal	Cancer	Pneumonia	COVID
Normal	435	6	4	5
Cancer	8	430	7	5
Pneumonia	5	9	428	8
COVID	4	6	10	430

The confusion matrix (Table 8) shows that the proposed model accurately distinguishes between different lung diseases. The majority of classification errors were between pneumonia and COVID-19, which have similar radiological appearances.

4.6 ROC Curve Analysis

The classification performance of each class is good, as shown by the ROC curves. The total across all classes was over 0.95, a strength of the proposed model.

Table 9. *Class-wise AUC Scores*

Disease Class	AUC
Normal	0.98
Lung Cancer	0.97
Pneumonia	0.96
COVID-19	0.97

Table 9 shows a high AUC across all disease classes, indicating high classification performance. The model has the highest AUC for Normal (0.98), then Lung Cancer and COVID-19 (0.97), and Pneumonia (0.96), indicating strong discriminative ability across all categories.

4.7 Comparison with Existing Models

To demonstrate the usefulness of the suggested method, it was compared with several popular deep learning models used in medical image classification.

Table 10. *Model Performance Comparison*

Model	Accuracy	Precision	Recall	F1 Score
VGG16	90.8%	90.2%	89.7 %	89.9 %
ResNet50	91.4%	91.0%	90.6 %	90.8 %
DenseNet121	92.6%	92.1%	91.8 %	91.9 %
Vision Transformer	93.1%	92.8%	92.5 %	92.6 %
Attention	94.2%	93.7%	93.4	93.5

CNN			%	%
Proposed Hybrid CNN-Transformer	**96.3%**	**95.9%**	**95.8 %**	**95.8 %**

The results of various deep learning models in classification are compared in Table 10. The more advanced models, such as DenseNet121 and Vision Transformer, build on classic models, including VGG16 and ResNet50, and achieve 90-91% accuracy. Attention CNN also increases to 94.2%. The Hybrid CNN-Transformer model offers the best results, with 96.3% accuracy and higher precision, recall, and F1-score, indicating that it is more effective than other methods. The suggested hybrid architecture was much more efficient than traditional CNN models and single Transformer architectures.

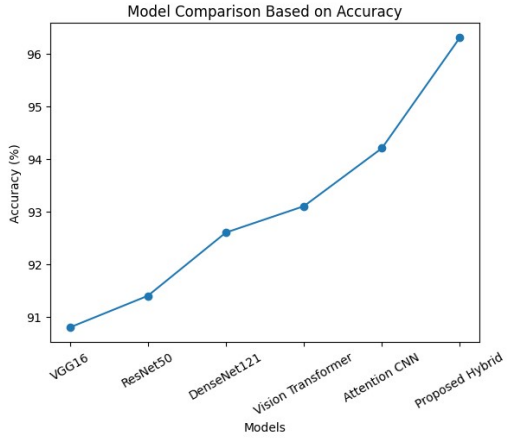

Figure 5. *Comparative Analysis of Deep Learning Models Based on Classification Accuracy*

Figure 5 presents the comparison of classification efficacies of different deep learning models. Classical designs like VGG16 and ResNet50 achieve 90% and 91% accuracy, respectively, and the most modern designs like DenseNet 121 and Vision Transformer achieve even higher levels of accuracy. Accuracy improves further to 94.2 with the Attention CNN. It is important to note that the proposed Hybrid CNNTransformer model achieves the highest accuracy of 96.3, surpassing that of the other models. It indicates that CNN-based feature extraction is effective when used with transformer-based contextual learning to enhance classification. Comparison of various deep learning models in terms of correctness in the classification. Conventional architectures like VGG16 and ResNet50 achieve accuracies of 90-91, while other models like DenseNet121 and Vision Transformer achieve higher accuracies. The CNN-Attention further improves the accuracy to 94.2. It is worth noting that the proposed Hybrid CNN-Transformer model achieves the highest accuracy of 96.3 percent, surpassing all other models. It shows that CNN-based feature extraction combined with transformer-based contextual learning has a better chance of improving classification performance.

4.8 Ablation Study

An ablation study was conducted to assess the value of each element of the presented framework.

Table 11. *Performance Comparison of Model Variants Based on Classification Accuracy*

Model Variant	Accuracy
CNN Only	91.7%
Transformer Only	92.8%
CNN + Transformer	94.5%
CNN + Transformer + Attention Fusion	**96.3%**

Table 11 shows a clear gain in accuracy as the model design becomes more advanced. Although CNN-only and Transformer-only models achieve moderate results, their combination achieves 94.5% accuracy. The maximum accuracy of 96.3 percent is achieved when the attention integration module is added, demonstrating the efficiency of the feature representation. The findings affirm that the Multi-Scale Attention Fusion component significantly enhances the hybrid model.

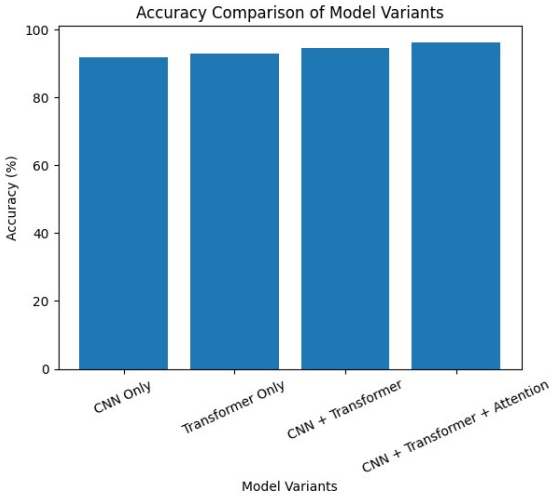

Figure 6. *Accuracy Comparison of Different Model Variants*

The performance of different model settings in terms of classification effectiveness is shown in Figure 6. The CNN-only and Transformer-only models achieve 91.7% and 92.8% accuracy, respectively. The combination of the two methods increases precision to 94.5, demonstrating the advantage of integrating spatial and contextual features. The complete model, including the attention fusion component, achieves 96.3 percent accuracy, indicating the effectiveness of multi-scale

attention in enhancing feature representation and overall performance.

4.9 Discussion

The experimental findings show that an effective combination of CNN and Transformer architectures has a strong positive influence on lung disease detection. The CNN component, which is good at identifying spatial features (nodules, lesions, and lung structure changes), and the Transformer component, which is good at identifying overall contextual relationships in CT images, complement each other. We have also added a Multi-Scale Attention Fusion module to combine features across multiple spatial resolutions, further enhancing results. The upgrade allows the model to recognize even the slightest localized abnormalities, not to mention the more serious pathology patterns in lung tissues. Compared with traditional CNN models such as VGG and ResNet, our model achieves an impressive 4-5% improvement in classification accuracy. Additionally, this hybrid design is significantly superior to unified designs based on Vision Transformer frameworks and demonstrates the advantage of combining the ability to extract spatial features with that to model contextual factors simultaneously. To conclude, our proposed framework for automated detection of lung diseases is a strong and effective decision-support tool for clinical diagnosis.

The study provides insights into the design of AI-powered medical imaging systems. Results indicate both local spatial characteristics and global context are vital for accurate lung disease classification. The CNN learns fine-grained information, such as nodules and texture, while the Transformer models distant relationships within the lung. The hybrid architecture shows that combining both perspectives yields a more complete feature representation.

The Multi-Scale Attention Fusion module is crucial for achieving these outcomes. Ablation study results show the attention mechanism extracts subtle disease features that traditional architectures often miss. This boosts model performance. This finding is especially important in medical image analysis, where fusing multiscale information helps distinguish diseases with similar radiological appearances.

The proposed model delivers performance improvements over state-of-the-art solutions. While the numerical improvement may appear small, it can be clinically significant. The system achieves 96.3% accuracy, surpassing DenseNet121, Vision Transformer, and Attention-CNN. Even a 2–5% increase in diagnostic screening accuracy could reduce missed cases and misdiagnoses, improving health outcomes and enabling earlier treatment. The results affirm complementary benefits from local feature extraction, contextual learning, and multi-scale attention fusion, which existing architectures lack. This demonstrates the potential of the proposed model to advance AI-based lung disease diagnosis.

The suggested system may help radiologists analyze CT images quickly and accurately. It is not intended to replace clinical experts. Instead, it serves as a decision-support tool to reduce diagnostic burden, improve consistency, and enable earlier diagnosis of lung diseases. The results show that hybrid deep learning systems will become increasingly important in future AI-based health care systems.

Despite these limitations, the proposed framework represents a significant step forward in applying advanced AI to lung disease diagnosis. Addressing the outlined challenges will be crucial for broader clinical adoption, but the model's strengths and substantial improvements over existing systems demonstrate its great potential to transform medical imaging and patient care. Continued refinements and future integrations of additional data sources and real-world scenarios will further strengthen the impact and reliability of such AI-powered systems.

4.10 Real-Life Implementation and Practical Applications

The proposed Hybrid CNN–Transformer approach can be applied to create computer-aided diagnostic (CAD) systems for lung disease screening and diagnosis in a healthcare radiology workflow. It automatically resolves the problems of lung disease diagnosis from a computed tomography (CT) scan image of a patient in a real-life application into possible classes, including lung cancer, pneumonia, COVID-19, or a normal lung condition. This could benefit radiologists by highlighting potentially suspicious cases, reducing the time doctors spend reviewing X-rays, and focusing their review of films on more difficult cases. This can be especially beneficial in high-volume healthcare settings where numerous CT scans need to be reviewed daily.

Moreover, the proposed framework can provide telemedicine or other telehealthcare services, with the initial diagnosis conducted in areas where radiology services are insufficient. The model could also be incorporated into clinical decision-making support systems, improving the

consistency of clinical decisions, enabling early disease detection, and, consequently, improving patient outcomes. If validated and approved for clinical use, the suggested approach could also be effective for diagnosing pulmonary diseases using artificial intelligence (AI) in clinical practice.

5. CONCLUSION

This study proposed a Hybrid CNN-Transformer model with Multi-Scale Attention Fusion to detect lung disease from CT scans automatically. The primary aim of the paper was to improve the diagnostic quality of automated methods by leveraging the most powerful local spatial feature detector in Convolutional Neural Networks and the general contextual learning of Transformer designs. The proposed methodology consisted of CNN-based layered feature extraction, transformer-based contextual modelling and a new multi-scale attention fusion technique to enhance the feature representation and classification accuracy.

The experimental findings demonstrated that the proposed model achieved a classification accuracy of 96.3%, precision of 95.9%, sensitivity (recall) of 95.8%, F1-score of 95.8%, specificity of 96.9%, and ROC-AUC of 0.97. Compared to typical deep learning models, such as VGG16 (90.8%), ResNet50 (91.4%), DenseNet121 (92.6%), and Vision Transformer (93.1%), the proposed hybrid model was found to be better (91.37% overall classification error was improved by between 3 and 5). These findings confirmed that CNN-based spatial feature extraction, combined with Transformer-based contextual representation, improved lung disease detection on CT images. Although this study yielded positive results, several limitations were identified. The model has been trained and tested on publicly available datasets, though these may not capture the diversity and variability of clinical settings. In addition, the architectures proposed in most of the research were 2D CT slice-based rather than full 3D volumetric CT, which could have impaired the stability of the architectures during spatial relationship measurements across slices. Furthermore, the computational complexity of hybrid CNN–Transformer architectures may pose challenges for deployment in resource-constrained environments. The directions for future research are to implement the proposed structure in 3D volumetric CT analysis, to apply explainable AI methods to improve the model's explainability, and to create lightweight models for clinical use within the near future. In addition, the proposed system will be validated in future research using large-scale, multi-institutional datasets to improve the model's stability and generalization.

A few limitations of the previous literature and the present study will be addressed in future research. Currently, the methods commonly used to detect lung disease, such as the proposed framework, rely solely on 2D CT slices rather than on the spatial relationships within 3D volumetric CT data. Hence, hybrid CNN–Transformer models could be used in the future for 3D volumetric analysis of CT scans for providing better context and diagnosis. Moreover, little effort has been devoted to making the model interpretable, which is crucial for clinical applications. Future studies on automated diagnostic systems should focus on implementing Explainable Artificial Intelligence (XAI) methods, such as attention visualization and saliency maps, to increase transparency and trust in the system. One of the follow-up research lines is to corroborate the proposed framework using larger, multi-institutional datasets collected from different patient groups and imaging devices, thereby making it more robust and generalizable.

In the future, a lightweight, economical hybrid architecture suitable for resource-limited healthcare environments and edge-based medical systems can be explored. Moreover, multimodal clinical data, such as patient history, laboratory results, and radiology reports, can be combined with CT image analysis to enhance diagnostic accuracy and facilitate more comprehensive clinical decision-making. The guidelines can complement previous research publications in the field and accelerate the advancement of AI-powered medical diagnosis of lung diseases.

The key novel aspects of this research are the integration of the proposed Multi-Scale Attention Fusion mechanism into CNN-based spatial feature extraction, Transformer-based contextual learning, and the training of the entire system within a single framework for multi-class lung disease classification from CT images. The proposed model architecture naturally combines local and global features and enhances feature representation through adaptive multi-scale fusion, in contrast with traditional CNN models, which typically use only local features, and standalone Transformer models, which mainly rely on global context. The performance of this design is compared with that of some well-known designs, and the experimental results are presented. The latter offers all the performance improvements and is a very powerful, scalable solution for diagnosing lung diseases, thereby advancing the evolution of AI-driven

medical imaging. The findings highlight the potential of a hybrid deep-learning platform for clinical decision support, diagnostic accuracy, and the construction of future next-generation computer-aided diagnostic systems.

Overall, the presented CNN-Transformer hybrid algorithm showed promise as a potential tool for early lung disease detection via AI, applicable to more effective medical imaging processing and helping radiologists in diagnostic efforts.

Also, the research reveals that adopting local feature learning, global contextual modeling, and multi-scale attention mechanisms is a promising strategy for medical image classification compared to a single deep learning paradigm. The techniques developed in this study can be applied to the design of future computer-aided diagnostic (CAD) systems for other medical imaging applications in which both spatial detail and contextual knowledge are significant.

REFERENCES

[1] S. Momtazmanesh *et al.*, "Global burden of chronic respiratory diseases and risk factors, 1990–2019: An update from the Global Burden of Disease Study 2019," *eClinicalMedicine*, vol. 59, p. 101936, 2023, doi: 10.1016/j.eclinm.2023.101936.

[2] Y. E. Yu and C. Li, "Global trends and attributable risk factors in the disease burden of lower respiratory infections," *Tropical Medicine and Infectious Disease*, vol. 10, no. 7, p. 180, 2025, doi: 10.3390/tropicalmed10070180.

[3] M. Guirado, E. F. Martín, A. F. Villar, A. N. Martín, and A. Sánchez-Hernández, "Clinical impact of delays in the management of lung cancer patients in the last decade: Systematic review," *Clinical and Translational Oncology*, vol. 24, 2022, doi: 10.1007/s12094-022-02796-w.

[4] A. Nathani and H. E. Dincer, "Advancements in imaging technologies for the diagnosis of lung cancer and other pulmonary diseases," *Diagnostics*, vol. 15, no. 7, p. 826, Mar. 2025, doi: 10.3390/diagnostics15070826.

[5] K. Irwin, "CT scans significantly more effective than chest X-rays in reducing lung cancer deaths," *UCLA Health*, Jun. 29, 2011. [Online]. Available: https://www.uclahealth.org/news/release/ct-scans-significantly-more-effective-than-chest-x-rays

[6] H. Zamani, T. Fruscello, J. Burleson, M. Bhargavan-Chatfield, and M. S. Davenport, "US radiology imaging and workforce volumes 2017–2024: An analysis of 46.4 million imaging examinations from 167 radiology facilities," *J. Am. Coll. Radiol.*, early access, Dec. 26, 2025, doi: 10.1016/j.jacr.2025.12.026.

[7] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in CT," *European Radiology*, vol. 29, no. 3, pp. 1391–1399, Mar. 2019, doi: 10.1007/s00330-018-5695-5.

[8] A. I. S. Ahmad, J. Dai, Y. Xie, and X. Liang, "Deep learning models for CT image classification: A comprehensive literature review," *Quantitative Imaging in Medicine and Surgery*, vol. 15, no. 1, pp. 962–1011, Jan. 2025, doi: 10.21037/qims-24-1400.

[9] C. Dubois *et al.*, "Deep learning in medical image analysis: Introduction to underlying principles and reviewer guide using diagnostic case studies in paediatrics," *BMJ*, vol. 387, p. e076703, 2024, doi: 10.1136/bmj-2023-076703.

[10] Y. Bian, J. Li, C. Ye, X. Jia, and Q. Yang, "Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models," *Chinese Medical Journal*, vol. 138, no. 6, pp. 651–663, Mar. 2025, doi: 10.1097/CM9.0000000000003489.

[11] P. Muthukumaraswamy, T. Yuvaraj, and R. Krishnamoorthy, "Semi-supervised multi-class pneumonia classification using a CNN-cascade forest framework," *Scientific Reports*, vol. 16, no. 1, p. 7448, Feb. 2026, doi: 10.1038/s41598-026-38849-1.

[12] Z. Ur Rehman, Y. Qiang, L. Wang, Y. Shi, Q. Yang, S. U. Khattak, R. Aftab, and J. Zhao, "Effective lung nodule detection using deep CNN with dual attention mechanisms," *Scientific Reports*, vol. 14, no. 1, p. 3934, Feb. 2024, doi: 10.1038/s41598-024-51833-x.

[13] B. Liu, L. Wan, and G. Zhang, "Hybrid transformer-CNN framework for precise segmentation and classification of pulmonary nodules in CT images," *Journal of Radiation Research and Applied Sciences*, vol. 19, no. 1, p. 102259, 2026, doi: 10.1016/j.jrras.2026.102259.

[14] B. A., M. Kaur, D. Singh, S. Roy, and M. Amoon, "Efficient skip connections-based residual network (ESRNet) for brain tumor classification," *Diagnostics*, vol. 13, no. 20, p. 3234, Oct. 2023, doi: 10.3390/diagnostics13203234.

[15] G. Huang, Z. Liu, G. Pleiss, L. van der Maaten, and K. Q. Weinberger, "Convolutional networks with dense connectivity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8704–8716, Dec. 2022, doi: 10.1109/TPAMI.2019.2918284.

[16] C. Chen, N. A. Mat Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," *Computers in Biology and Medicine*, vol. 185, p. 109507, 2025, doi: 10.1016/j.compbiomed.2024.109507.

[17] S. Muksimova, S. Umirzakova, N. Iskhakova, A. Khaitov, and Y. I. Cho, "Advanced convolutional neural network with attention mechanism for Alzheimer's disease classification using MRI," *Computers in Biology and Medicine*, vol. 190, p. 110095, 2025, doi: 10.1016/j.compbiomed.2025.110095.

[18] S. R. Choi and M. Lee, "Transformer architecture and attention mechanisms in genome data analysis: A comprehensive review," *Biology (Basel)*, vol. 12, no. 7, p. 1033, Jul. 2023, doi: 10.3390/biology12071033.

[19] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, "Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 6, pp. 3928–3971, Dec. 2025, doi: 10.1007/s10278-025-01481-y.

[20] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives," *Medical Image Analysis*, vol. 85, p. 102762, Apr. 2023, doi: 10.1016/j.media.2023.102762.

[21] B. Xu, J. Yan, H. Tang, Y. Zhang, M. Wang, and J. Gao, "An efficient compressive CNN and transformer hybrid framework for long-term dissolved oxygen prediction in aquaculture," Information Processing in Agriculture, 2026, doi: 10.1016/j.inpa.2026.01.011.

[22] J. W. Kim, A. U. Khan, and I. Banerjee, "Systematic review of hybrid vision transformer architectures for radiological image analysis," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 5, pp. 3248–3262, Oct. 2025, doi: 10.1007/s10278-024-01322-4.

[23] S. L. Tan, G. Selvachandran, W. Ding, and K. Kotecha, "Artificial intelligence in lung cancer imaging: A review of framework architectures and computer-aided diagnosis advancements," *Engineering Applications of Artificial Intelligence*, vol. 173, p. 114481, 2026, doi: 10.1016/j.engappai.2026.114481.

[24] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: A review," *arXiv preprint*, 2017, doi: 10.48550/arXiv.1709.02250.

[25] R. Javed, T. Abbas, A. H. Khan, A. Daud, A. Bukhari, and R. Alharbey, "Deep learning for lung cancer detection: A review," *Artificial Intelligence Review*, 2024, doi: 10.1007/s10462-024-10807-1.

[26] Y. Liu, Y. Liu, Y. Liu, and Y. Liu, "Deep learning-based lung cancer classification of CT images," *BMC Cancer*, 2025, doi: 10.1186/s12885-025-14320-8.

[27] B. Jiang, N. Li, X. Shi, S. Zhang, J. Li, G. H. Bock, R. Vliegenthart, and X. Xie, "Deep learning reconstruction shows better lung nodule detection for ultra-low-dose chest CT," *Radiology*, vol. 303, no. 1, pp. 202–212, 2022, doi: 10.1148/radiol.210551.

[28] K. Abdullahi, K. Ramakrishnan, and A. B. Ali, "Deep learning techniques for lung cancer diagnosis with computed tomography imaging: A systematic review for detection, segmentation, and classification," Information, vol. 16, no. 6, 2025, doi: 10.3390/info16060451.

[29] S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A survey of deep learning for lung disease detection on medical images: State-of-the-art, taxonomy, issues and future directions," *Journal of Imaging*, vol. 6, no. 12, p. 131, 2020, doi: 10.3390/jimaging6120131.

[30] H. Jung, B. Kim, I. Lee, J. Lee, and J. Kang, "Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method," *BMC Medical Imaging*, vol. 18, no. 1, p. 48, 2018, doi: 10.1186/s12880-018-0286-0.

[31] A. Raza, F. Hanif, and H. A. Mohammed, "Clinical validation of lightweight CNN architectures for reliable multi-class classification of lung cancer using histopathological imaging techniques,"

Scientific Reports, vol. 16, no. 1, p. 6512, Jan. 2026, doi: 10.1038/s41598-026-36652-6.

[32] M. K. Faizi, Y. Qiang, Y. Wei, Y. Qiao, J. Zhao, R. Aftab, and Z. Urrehman, "Deep learning-based lung cancer classification of CT images," *BMC Cancer*, vol. 25, no. 1, p. 1056, Jul. 2025, doi: 10.1186/s12885-025-14320-8.

[33] G. Wang, L. Cheng, J. Lin, Y. Dai, and T. Zhang, "Fine-grained classification based on multi-scale pyramid convolution networks," *PLoS One*, vol. 16, no. 7, p. e0254054, 2021, doi: 10.1371/journal.pone.0254054.

[34] H. Zhang, Y. Peng, and Y. Guo, "Pulmonary nodules detection based on multi-scale attention networks," *Scientific Reports*, vol. 12, 2022, doi: 10.1038/s41598-022-05372-y.

[35] Y. Li, L. Hui, X. Wang, *et al.*, "Lung nodule detection using a multi-scale convolutional neural network and global channel spatial attention mechanisms," *Scientific Reports*, vol. 15, Art. no. 12313, 2025, doi: 10.1038/s41598-025-97187-w.

[36] B. Oltu, S. Güney, S. E. Yüksel, and B. Dengiz, "Automated classification of chest X-rays: A deep learning approach with attention mechanisms," *BMC Medical Imaging*, vol. 25, no. 1, p. 71, 2025, doi: 10.1186/s12880-025-01604-5.

[37] Z. Ullah, M. Hong, T. Mahmood, and J. Kim, "Systematic integration of attention modules into CNNs for accurate and generalizable medical image classification," *Mathematics*, vol. 13, no. 22, 2025, doi: 10.3390/math13223728.

[38] W. Liu, J. Sun, H. Li, Y. Wang, and Z. Wang, "CSEA-Net: A channel-spatial enhanced attention network for lung tumor segmentation on CT images," *iScience*, vol. 28, no. 3, 2025, doi: 10.1016/j.isci.2025.111974.

[39] G. Liu, F. Liu, J. Gu, X. Mao, X. Xie, and J. Sang, "An attention-based deep learning network for lung nodule malignancy discrimination," *Frontiers in Neuroscience*, vol. 16, 2023, doi: 10.3389/fnins.2022.1106937.

[40] S. Tiwari, A. Shukla, and A. K. Sharma, "Attention-guided lightweight deep learning architecture for classification of polycystic ovary syndrome from ultrasound images," *Intelligence-Based Medicine*, vol. 13, p. 100358, 2026, doi: 10.1016/j.ibmed.2026.100358.

[41] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," *arXiv preprint arXiv:2206.09959*, 2022, doi: 10.48550/arXiv.2206.09959.

[42] W. Sun, Y. Pang, and G. Zhang, "CCT: Lightweight compact convolutional transformer for lung disease CT image classification," *Frontiers in Physiology*, vol. 13, 2022, doi: 10.3389/fphys.2022.1066999.

[43] J. W. Kim, A. U. Khan, and I. Banerjee, "Systematic review of hybrid vision transformer architectures for radiological image analysis," *Journal of Imaging Informatics in Medicine*, vol. 38, 2025, doi: 10.1007/s10278-024-01322-4.

[44] Zhang, Y., Wang, J., Gorriz, J. M. & Wang, S. (2023). Deep Learning and Vision Transformer for Medical Image Analysis. J. Imaging 2023. https://doi.org/10.3390/jimaging9070147

[45] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin, B. Silva, J. Sena, B. Shickel, A. Bihorac, K. Khezeli, and P. Rashidi, "Transformers and large language models in healthcare: A review," *Artificial Intelligence in Medicine*, vol. 154, p. 102900, 2024, doi: 10.1016/j.artmed.2024.102900.

[46] X. Liu, Y. Hu, and J. Chen, "Hybrid CNN-Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron," *Biomedical Signal Processing and Control*, vol. 86, part C, Art. no. 105331, 2023, doi: 10.1016/j.bspc.2023.105331.

[47] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From CNN to Transformer: A review of medical image segmentation models," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 4, pp. 1529–1547, Aug. 2024, doi: 10.1007/s10278-024-00981-7.

[48] S. Rezvani, M. Fateh, Y. Jalali, and A. Fateh, "FusionLungNet: Multi-scale fusion convolution with refinement network for lung CT image segmentation," *arXiv preprint*, 2024. doi: 10.48550/arXiv.2410.15812.

[49] Y. Huang, T. Wang, S. Huang, J. Zhang, H. Chen, Y. Chang, and R. Chang, "An improved 3-D attention CNN with hybrid loss and feature fusion for pulmonary nodule classification," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107278, Feb. 2023, doi: 10.1016/j.cmpb.2022.107278.

[50] W. Zhu, C. Liu, W. Fan, and X. Xie, "DeepLung: 3D deep convolutional nets for automated pulmonary nodule detection and classification," *arXiv preprint*

arXiv:1709.05538, 2017. doi: 10.48550/arXiv.1709.05538.

[51] Y. Gu, X. Lu, L. Yang, B. Zhang, D. Yu, Y. Zhao, L. Gao, L. Wu, and T. Zhou, "Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs," *Computers in Biology and Medicine*, vol. 103, pp. 220–231, 2018, doi: 10.1016/j.compbiomed.2018.10.011.

[52] X. Fu, R. Lin, W. Du, A. Tavares, and Y. Liang, "Explainable hybrid transformer for multi-classification of lung disease using chest X-rays," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-90607-x.

[53] MosMedData, "MosMedData: Chest CT Scans with COVID-19 Related Findings," 2020. [Online]. Available: https://mosmed.ai/datasets/covid19_1110

[54] COVID-CT Dataset, "COVID-CT Dataset," GitHub Repository, 2020. [Online]. Available: https://github.com/UCSD-AI4H/COVID-CT

[55] LIDC-IDRI, "LIDC-IDRI Dataset," The Cancer Imaging Archive, 2011. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI