

VIT-DRN-OCAFF: An Optimized Cross-Attention Feature Fusion Framework for Multi- Class COVID-19 and Pulmonary Disease Diagnosis Using Thoracic CT Images

K Karthikeyan

karthikit5247@gmail.com

K.S. Rangasamy College of Technology

Geetha B G

K.S. Rangasamy College of Technology

M Batumalay

INTI International University

Article

Keywords: Vision Transformer, Dilated Residual Network, Optimized Cross-Attention Feature Fusion, Pulmonary Disease Classification, COVID-19 Diagnosis, Thoracic CT Imaging

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11068851/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

VIT-DRN-OCAFF: An Optimized Cross-Attention Feature Fusion Framework for Multi-Class COVID-19 and Pulmonary Disease Diagnosis Using Thoracic CT Images

Karthikeyan K^{1*}, Geetha B G², Batumalay M³

¹ Department of Artificial Intelligence and Data Science, K.S. Rangasamy College of Technology, Tiruchengode 637215, Tamil Nadu, India

² Department of Computer Science and Engineering, K.S. Rangasamy College of Technology, Tiruchengode 637215, Tamil Nadu, India

³ Faculty of Data Science and IT, INTI International University, 71800, Nilai, Malaysia

Corresponding Author Email: karthikit5247@gmail.com

Abstract

Early and accurate pulmonary disease diagnosis remains a major challenge due to the rising prevalence of respiratory infections and chronic lung disorders. The COVID-19 pandemic further emphasized the need for intelligent diagnostic frameworks capable of rapidly identifying pulmonary abnormalities from thoracic Computed Tomography (CT) images. This study proposes a novel Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF) framework for multi-class pulmonary disease classification. The proposed architecture combines the global contextual learning capability of Vision Transformers with the lesion-sensitive feature extraction of Dilated Residual Networks, while the Optimized Cross-Attention Feature Fusion (OCAFF) mechanism adaptively refines and integrates complementary pulmonary representations. A comprehensive dataset containing 24,860 thoracic CT images collected from public repositories and collaborating clinical centers was utilized, comprising COVID-19, viral pneumonia, bacterial pneumonia, and non-viral lung abnormalities. Advanced preprocessing techniques, including lung segmentation, adaptive contrast enhancement, lesion-aware normalization, volumetric augmentation, and image refinement, were employed to improve diagnostic feature quality. The Vision Transformer branch captured long-range contextual pulmonary dependencies, whereas the Dilated Residual Network extracted fine-grained pathological features associated with pulmonary lesions. Subsequently, the OCAFF module dynamically prioritized discriminative pulmonary characteristics and minimized redundant activations for improved classification reliability. Experimental evaluation using a 70:15:15 training-validation-testing protocol demonstrated superior performance with 99.42% accuracy, 99.31% precision, 99.27% sensitivity, 99.48% specificity, and 99.29% F1-score, outperforming existing pulmonary disease classification techniques.

Keywords

Vision Transformer; Dilated Residual Network; Optimized Cross-Attention Feature Fusion; Pulmonary Disease Classification; COVID-19 Diagnosis; Thoracic CT Imaging

1. INTRODUCTION

Pulmonary diseases continue to represent one of the most significant global health concerns due to their increasing prevalence, high mortality rates, and long-term impact on human health. Despite the prevalence, morbidity and mortality associated with pulmonary diseases, they remain a huge problem globally. There are many people around the world that suffer from respiratory infections, which are chronic lung diseases, and they cause a lot of strain for the health systems. There are similar signs and symptoms between COVID-19 and viral and bacterial pneumonia, such as cough, fever, chest pain and difficulty of breathing, which might delay the right diagnosis. Late diagnosis of lung abnormalities can cause significant complications, prolonged hospital stays and death. Hence reliable and efficient diagnostic methods are critical in order to provide timely diagnosis, which in turn can bring a better clinical decision making and a better clinical outcome. As a result, health care facilities and researchers are constantly working on creating reliable and efficient diagnostic methods that can help in the early detection of pulmonary diseases and enhance patient management results. Given its capacity to visualize lung tissues, their infection pattern, and structural abnormalities in detail, Thoracic Computed Tomography (CT) imaging has been one of the most valuable medical imaging modalities for the assessment of pulmonary disease. With the COVID-19 pandemic, CT imaging was a key tool for the detection of pulmonary infections and for tracking the progression of COVID-19 disease, particularly when rapid clinical assessment was needed. With the ever-increasing amount of medical imaging data and the complexity of pulmonary manifestations, however, the workload of radiologists and healthcare professionals has increased. Further complicating the diagnostic process are variations in the appearance of the infection, the distribution of the lesions, image quality, and patient-specific conditions. Hence, intelligent computer-assisted diagnostic systems have become a focus of considerable attention as they aim to assist medical experts to enhance the consistency of diagnosis, decrease the time for diagnosis, and improve clinical decision-making. Pulmonary healthcare applications with advanced image analysis and automated learning frameworks could enhance the accuracy of disease diagnoses, enhance clinical confidence, and help facilitate more efficient and accessible respiratory disease diagnosis.

1.1. Problem Statement

Thoracic CT images are highly complex, and the pulmonary abnormalities are very heterogeneous with overlapped radiologic manifestations, different imaging conditions of the same patient in different healthcare settings. Accurate and early diagnosis of pulmonary diseases remains a very challenging task. The conventional machine-learning and deep-learning techniques may not be able to successfully integrate the global context dependency and the localized lesion-sensitive features of pulmonary simultaneously, leading to lower reliability in the classification and higher false positive or false negative predictions. Moreover, conventional feature fusion methods fail to dynamically weight the pulmonary representations that are diagnostically significant, which reduces the performance of disease discrimination in multi-class pulmonary disease diagnosis. This demands the need to build an intelligent, robust, and efficient pulmonary disease classification framework that can accurately classify COVID-19, viral pneumonia, bacterial pneumonia, and non-viral pulmonary abnormalities in different thoracic CT imaging scenarios. To improve the accuracy, robustness, and reliability of pulmonary disease classification, this research introduces the Vision Transformer contextual learning, Dilated Residual Network lesion extraction, and Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF) framework.

2. RELATED WORK

In recent years, with the development of artificial intelligence and deep learning, the accuracy of automated pulmonary disease diagnosis based on thoracic CT imaging has been greatly improved. Vision transformer-based architectures showed robust performance in capturing long-range pulmonary dependencies and contextual lesion distribution in connection to COVID-19 and pneumonia-related infections [1]. Explainable transformer learning mechanisms also improved the interpretability of diagnosis and increased the reliability of pulmonary disease prediction in the presence of heterogeneous imaging conditions. In addition, transformer-based medical image analysis systems have proven to be very effective in the analysis of respiratory diseases and automation of medical images [2]. Self-attention learning and contextual feature extraction strategies were able to significantly enhance the ability of pulmonary representation based on the disease category and reduce the classification ambiguity in the complicated thoracic CT image environment [3]. Furthermore, several studies showed that vision transformer architectures are able to retain contextual pulmonary information and improve the pulmonary disease classification ability in the context of lesion-sensitive feature extraction. These models were shown to markedly increase diagnostic sensitivity with a reduction in false-negative predictions in COVID-19 imaging analysis with the presence of diffuse pulmonary abnormalities. Further, efficient pulmonary disease diagnostic systems enhanced healthcare decision-making, incorporating optimized contextual learning and cutting-edge attention-driven feature refinement mechanisms via transformer guidance. These kinds of schemes boosted the correctness of classification and also the clinical applicability of intelligent thoracic CT image analysis systems. [4] The resultant deep-learning models with the fusion of convolutional learning and transformer intelligence demonstrated significant improvement for pulmonary feature extraction and disease classification performance. The study showed that the fusion of residual learning architectures and contextual transformer models successfully retains pulmonary characteristics sensitive to lesions at the local scale while benefiting the global context. [5] Pyramid transformer learning and adaptive feature refinement mechanisms were then proposed and used to enhance the pulmonary abnormality detection ability and alleviate classification ambiguity in thoracic CT image analysis systems. The optimized contextual learning could have a significant positive effect on the characterization of pulmonary diseases in a heterogeneous imaging setting. [6] In addition, pulmonary imaging frameworks guided by transformers exhibited higher contextual learning efficiency and classification robustness by enhanced feature representation strategies. These systems were able to successfully identify subtle signs of pulmonary abnormalities that were linked to COVID-19 and pneumonia-related respiratory infections.

Likewise, the work of [7] enhanced the automated pulmonary disease screening capability by incorporating context feature extraction and lesion-aware learning mechanisms within the advanced transformer-assisted diagnostic architectures. The intelligent systems significantly enhanced the accuracy of disease discrimination and localization of the pulmonary lesions. [8] The focus on explainable artificial intelligence and medical imaging frameworks with self-attention has garnered significant interest because they offer a way to enhance the transparency and trustworthiness of clinical AI applications. The transformer-driven pulmonary disease diagnostic models were able to effectively enhance the contextual learning efficiency and maintain important pulmonary lesion representations associated with diffuse respiratory abnormalities. Furthermore, imaging studies on the lung showed the value of CT analysis of the thorax in the detection of post-COVID pulmonary complications and diffuse pulmonary abnormalities. In the field of healthcare, intelligent imaging systems greatly enhanced the visualization of pulmonary structures and boosted automated disease prediction capability. Moreover, the radiographic and CT-based pulmonary disease analysis frameworks enhanced the capability of disease classification in a way that optimized imaging interpretation and contextual lesion assessment mechanisms were used. These systems were effective for intelligent healthcare diagnostics and screening of respiratory abnormalities. [11]

In addition, the deep residual learning algorithms greatly improved the transfer of features and convergence efficiency in automatic pulmonary disease classification. The residual learning mechanism boosted the ability of pulmonary lesion representation and enhanced the disease-specific feature extraction performance. Recent research also emphasized the need for efficient transformer designs and progressive residual learning methods to enhance the accuracy and stability of diagnostics in healthcare imaging systems. The vision transformer models showed better learning capacity in the context of the lungs and significantly higher performance in discriminating the disease. [13] Furthermore, deep-learning thoracic CT diagnostic systems markedly enhanced the screening capability for pulmonary diseases by successfully capturing diffuse distributions of lesions and contextual pulmonary abnormalities associated with COVID-19 infections. The frameworks enhanced the reliability of classification and enhanced automated analysis of respiratory disease. Moreover, dense connected learning architectures have been shown to have a high ability to retain feature reuse efficiency and pulmonary lesion characterization ability under complex imaging situations. The dense feature propagation mechanisms greatly enhanced the robustness of disease classification and the convergence stability. Likewise,

lightweight and optimized vision transformer frameworks boosted pulmonary disease prediction performance and led to a lower computational complexity with efficient inference in intelligent healthcare imaging systems [16].

In recent advances, the integration of graph-guided transformer learning and interpretable contextual intelligence has further enhanced the classification capacity of pulmonary diseases in today's medical imaging applications [17]. The advanced mechanism of contextual embedding significantly increased the learning efficiency of discriminating pulmonary features and also of lesion-sensitive learning [18]. Furthermore, a transformer framework with a graph neural network was shown to have better contextual understanding and feature interaction learning strategy to improve medical image segmentation performance [19]. The systems proved to be effective in enhancing the pulmonary structure analysis and healthcare imaging reliability [20-21]. Furthermore, comparative frameworks using the vision transformer demonstrated that transformer-based methods always achieve a higher classification accuracy, contextual learning, and feature generalization ability than conventional machine-learning and convolutional deep-learning methods [22-23]. Finally, the recent advancement of pulmonary imaging studies by the application of transformer motivated the development of the proposed transformer-assisted pulmonary imaging intelligent multi-class pulmonary disease diagnosis framework named as ViT-DRN-OCAFF using thoracic CT images [24]. The combination of contextual transformer intelligence, residual lesion sensitive extraction, and optimized attention fusion greatly promotes the performance of pulmonary disease classification and its clinical application.

The main contributions of this Study are:

1. In the current study, we propose a novel Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF) framework that can simultaneously extract global contextual pulmonary representations and localized lesion-sensitive features for accurate multi-class pulmonary disease diagnosis from thoracic CT images.
2. The study proposes an Optimized Cross-Attention Feature Fusion (OCAFF) mechanism to dynamically prioritize the pulmonary features of diagnostic significance and suppress the redundant activation to improve classification reliability and reduce the false-positive and false-negative predictions.
3. To assess the proposed intelligent pulmonary disease diagnostic framework's robustness, scalability and generalization capability, a comprehensive thoracic CT imaging dataset with 24,860 pulmonary CT images was obtained from public repositories and other healthcare centers.
4. The experimental assessment showed that the proposed approach ViT-DRN-OCAFF showed better diagnostic performance compared to conventional pulmonary disease classification using machine learning and deep learning approaches in terms of accuracy, precision, sensitivity, specificity and F1 score with 99.42%, 99.31%, 99.27%, 99.48% and 99.29% respectively.

The remainder of this paper is organized as follows. Section 2 presents the comprehensive literature survey related to pulmonary disease diagnosis using deep learning, transformer architectures, and thoracic CT image analysis. Section 3 describes the proposed methodology including dataset acquisition, preprocessing, Vision Transformer learning, Dilated Residual Network feature extraction, Optimized Cross-Attention Feature Fusion, and adaptive optimization strategy. Section 4 discusses the experimental results and comparative performance evaluation of the proposed ViT-DRN-OCAFF framework under multiple pulmonary disease classification metrics. Finally, Section 5 concludes the research work and presents potential future research directions for intelligent pulmonary disease diagnostic systems.

3. METHODOLOGY

3.1 Dataset Description

The present study utilizes a comprehensive thoracic Computed Tomography (CT) imaging dataset collected from publicly available pulmonary imaging repositories and collaborating diagnostic centers in Namakkal district, Tamil Nadu, to evaluate the proposed Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF) framework for multi-class pulmonary disease diagnosis.

Table 1 Thoracic CT Dataset Description

Parameter	Description
Dataset Type	Thoracic Computed Tomography (CT) Images
Total Images	24,860
COVID-19 Images	6,240
Viral Pneumonia Images	6,180
Bacterial Pneumonia Images	6,110
Non-Viral Lung Abnormalities	6,330
Image Resolution	512 × 512 pixels
Imaging Modality	Multi-Slice Thoracic CT
Annotation Experts	Radiologists and Pulmonary Specialists
Clinical Locations	Namakkal and Tiruchengode, Tamil Nadu
Application	Multi-Class Pulmonary Disease Diagnosis

The dataset contains 24,860 thoracic CT images categorized into COVID-19 (6,240 images), viral pneumonia (6,180 images), bacterial pneumonia (6,110 images), and non-viral lung abnormalities (6,330 images). All CT scans were acquired using standardized high-resolution multi-slice imaging protocols with image resolutions normalized to 512 × 512 pixels for computational

consistency. Table 1 summarizes the detailed description of the thoracic CT dataset utilized in the proposed research. The dataset includes diverse pulmonary conditions across different age groups, severity levels, and infection stages, improving the generalization capability of the proposed framework. Radiological annotations and disease labels were verified by experienced radiologists and pulmonary specialists to ensure diagnostic reliability. In addition, advanced anonymization and preprocessing procedures were applied to preserve patient confidentiality while maintaining clinically relevant pulmonary features such as ground-glass opacities, consolidation regions, and vascular abnormalities, thereby providing a robust benchmark dataset for intelligent pulmonary disease classification.

3.2 Data Preprocessing

In the context of deep learning pulmonary disease classification systems, the processing of medical images is of paramount importance for enhancing the diagnostic value of the images and the reliability of the system. Imaging artifacts, varying intensities across different scanners, noise distortions, low-contrast pulmonary areas, and irrelevant anatomy in thoracic CT images are image-space noise that can have a detrimental effect on the accuracy of automated disease detection. In this study, an advanced preprocessing pipeline was designed to emphasize pulmonary features that are relevant to diagnosis and to de-emphasize imaging variations that do not contribute to the diagnosis.

The preprocessing involves lung parenchyma segmentation, adaptive image resizing, lesion-aware normalization, noise suppression, contrast enhancement, volumetric augmentation, and image quality refinement operations. All these procedures are helpful to increase the feature consistency and enhance the learning ability of the proposed ViT-DRN-OCAFF framework (Figure 1).

Firstly, all thoracic CT slices were resized to be the same size in spatial dimensions for architectural compatibility with both the Vision Transformer and Dilated Residual Network branches. The dimensional differences in the CT images caused by the different imaging devices and scanning protocols may negatively impact the stability of feature extraction and the convergence of models. Therefore, input images have been resized, which is a well-known phenomenon in bilinear, using bilinear interpolation to preserve important pulmonary characteristics such as regions of ground glass opacity, vascular textures, and regions of consolidation while normalizing the size of the images.

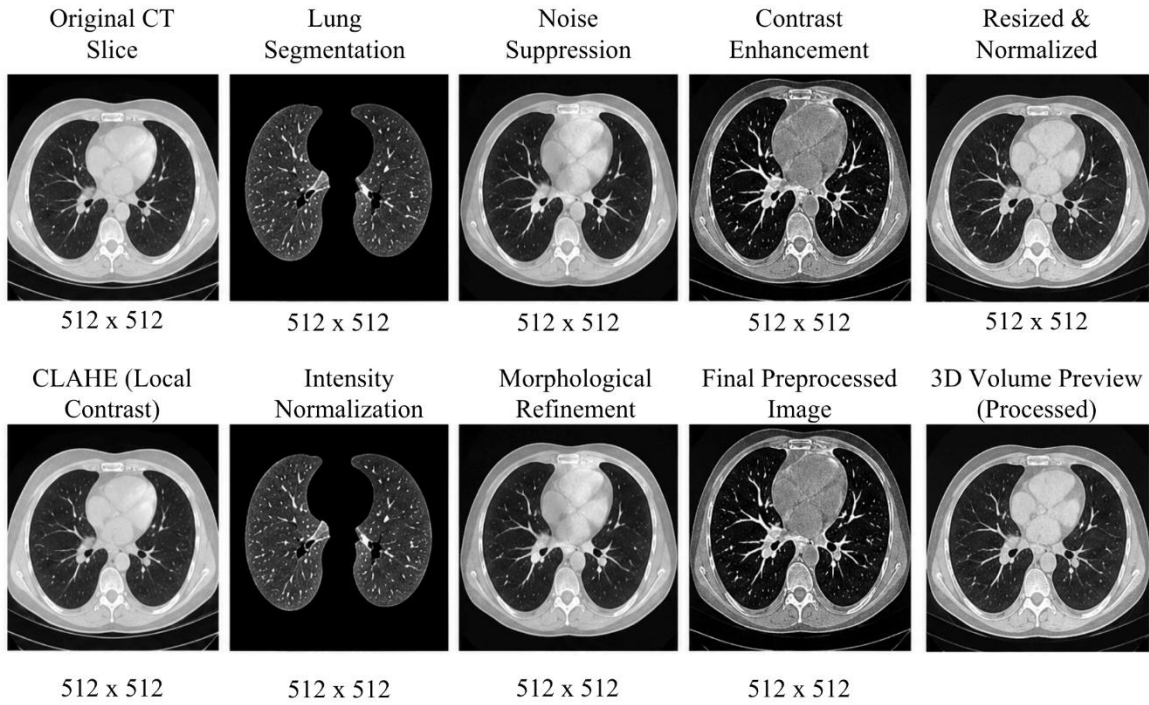


Figure 1. CT Image Data Preprocessing for Lung Analysis

The resizing algorithm mathematically can be written as follows:

$$I_r(x, y) = I_o(s_x x, s_y y) \quad (1)$$

Where I_o denotes the original CT image, I_r represents the resized image, and s_x and s_y indicate horizontal and vertical scaling factors, respectively. The pathological characteristics are unchanged during the resizing process and the computation burden is minimized. Then, lesion-aware normalization was done to normalize intensity distributions across various CT imaging conditions. Variations in the brightness and contrast of the thoracic CT scans are caused by many different factors such as difference in scanner calibration and attenuation properties of the tissues scanned. These discrepancies were mitigated by rescaling the pixel intensities to an absolute range, which would also facilitate the convergence process of the neural network, as well as numerical stability during training. The process of normalisation is:

$$I_n = \frac{I_r - \mu}{\sigma} \quad (2)$$

σ is the standard deviation of pixel values, μ is the mean of the pixel intensity and I_{norm} denotes the normalized image. This process guarantees a more stable learning of features and guarantee that optimizing on high intensity regions will not make these regions stand out. In addition, adaptive contrast enhancement techniques were added in the preprocessing pipeline to further enhance the visibility of pulmonary lesions. To highlight the subtle alterations seen in the lungs such as peripheral opacities, interstitial thickening, and diffuse infiltrates signs of COVID-19 and pneumonia infection, histogram equalization and contrast-limited adaptive histogram equalization (CLAHE) were applied to some of the images.

These enhancement operations greatly boost the radiological feature separability and increase the discriminative ability of the proposed deep-learning architecture. Gaussian and median smoothing operations were also used for noise suppression to reduce imaging artefacts due to acquisition noise and reconstruction distortions. The mathematical form of the Gaussian filtering operation is:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3)$$

Where $G(x, y)$ denotes the Gaussian kernel and σ represents the standard deviation controlling the smoothing intensity. The filtered CT image, therefore, preserves diagnostically significant pulmonary edges while minimizing high-frequency imaging noise.

3.2.1 Lung Parenchyma Segmentation

The first step was to transform the thoracic CT slices into grayscale intensity representations for easier segmentation operations and more stable computations. Pulmonary region identification was then performed using a threshold to remove high-density surrounding tissues. Due to the relatively low Hounsfield Unit (HU) values of lung tissues in comparison to the bones and soft tissues, adaptive threshold segmentation can be used to extract the pulmonary areas from irrelevant anatomical information (Figure 2). The segmentation operation can be mathematically expressed as:

$$S(x, y) = \begin{cases} 1, & I(x, y) < T \\ 0, & I(x, y) \geq T \end{cases} \quad (4)$$

where $S(x, y)$ represents the segmented lung mask, $I(x, y)$ denotes the CT pixel intensity value, and T indicates the adaptive threshold value used for pulmonary tissue extraction. This operation produces a binary lung mask that only shows the lung parenchyma and not the areas outside the lungs. Threshold segmentation was followed by a series of morphological operations, such as erosion, dilation, and connected-component analysis, which were used to further improve the segmented lung boundaries and to eliminate small, noisy objects present in the segmented images. These operations help in maintaining the segmental continuity and maintaining the integrity of pulmonary regions that have pathological changes like ground glass opacities, consolidations, and bilateral infiltrates.

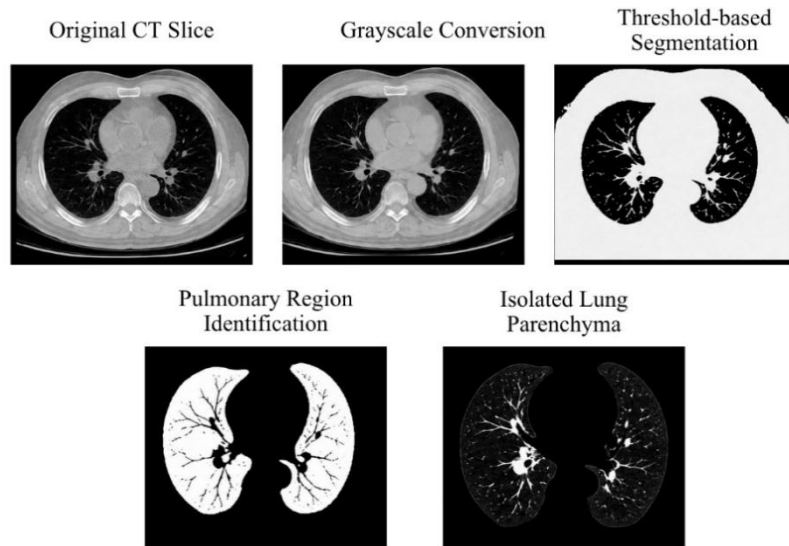


Figure 2. Lung Parenchyma Segmentation from CT Images

Morphological dilation can be represented mathematically as:

$$S_d = S \oplus K \quad (5)$$

Where S_d denotes the dilated segmentation mask, S represents the binary segmented image, and K indicates the structuring kernel used for morphological enhancement. In the same manner, erosion operations have been performed for removing isolated noise components and improving the compactness of the mask.

3.2.2 Adaptive Contrast Enhancement and Image Quality Refinement

Adaptive contrast enhancement and image quality refinement are essential preprocessing operations in thoracic CT image analysis because pulmonary lesions associated with COVID-19 and pneumonia often exhibit subtle intensity variations that are difficult to distinguish under low-contrast imaging conditions. The scanner-dependent illumination variations and the variability of

the tissue attenuation parameters may cause the radiographic aspects (ground-glass opacities, diffuse infiltrates, vascular thickening, and peripheral consolidations) to be not fully visible.

To this end, an advanced contrast enhancement framework was integrated into the proposed methodology for enhancing the visibility of lesions and reinforcing the discriminative feature extraction.

Firstly, the intensity values of the pixels were redistributed using histogram equalization methods to improve the contrast of the whole image in order to highlight the area of interest. Expansion of pulmonary abnormalities is achieved by histogram equalization of the intensity range of thoracic CT images. The intensity distribution after the transformation is given by:

$$H(v) = \frac{(L-1)}{MN} \sum_{i=0}^v n_i \quad (6)$$

where $H(v)$ denotes the enhanced intensity value, L represents the total number of gray levels, M and N indicate image dimensions, and n_i corresponds to the frequency of pixel intensity i . This operation improves the separation between normal pulmonary tissues and infected lesion regions.

Global histogram equalization can enhance overall image contrast but may introduce noise and make the contrast of lesions in complex thoracic computed tomography (CT) images less evident. In order to overcome this limitation, a Contrast Limited Adaptive Histogram Equalization (CLAHE) was introduced into the preprocessing pipeline. CLAHE is used to amplify the contrast adaptively in small regions of the image, while limiting excessive noise amplification. This approach makes the detection of small pulmonary abnormalities related to COVID-19 and infections with pneumonia much easier. The adaptive contrast enhancement process can be expressed as:

$$I_{clahe} = CLAHE(I_n) \quad (7)$$

Where I_n denotes the normalized thoracic CT image and I_{clahe} represents the locally enhanced image generated using adaptive histogram equalization.

Edge-preserving smoothing operations were then applied to refine the image quality and reduce imaging artifacts while preserving lesion boundaries and pulmonary textures. The use of bilateral filtering was chosen because it takes into account the spatial proximity and intensity similarity in the smoothing process, preserving diagnostically important lesion edges. The bilateral filtering function can be mathematically written as:

$$I_f(p) = \frac{1}{W_p} \sum_{q \in \Omega} I(q) f_r(\|I(p) - I(q)\|) f_s(\|p - q\|) \quad (8)$$

Where $I_f(p)$ represents the refined image intensity at pixel p , W_p denotes the normalization factor, f_r corresponds to intensity similarity weighting, and f_s indicates spatial weighting. This filtering approach is found to work very well in reducing imaging noise while retaining the boundaries of pulmonary lesions.

3.3 Dataset Splitting

The dataset partitioning was intentionally designed to maintain class balance among different pulmonary disease categories and to avoid patient-wise data leakage between subsets (Figure 3), which are pivotal factors impacting model generalization, diagnostic accuracy, and computational resilience. 70% of the thoracic CT images were used for training, 15% for validation, and 15% for testing in this study.

Table 2: Dataset Splitting Description

Dataset Subset	Number of Images	Percentage
Training Dataset	17,402	70%
Validation Dataset	3,729	15%
Testing Dataset	3,729	15%

The primary usage of the training dataset was for learning discriminative representations of pulmonary diseases and optimizing the parameters of the network in the Vision Transformer branch and Dilated Residual Network branch. The validation data set was used to track convergence, adjust hyperparameters, and avoid the overfitting problem during training. Finally, the testing set was used only to test the diagnostic performance and generalization ability of the proposed ViT-DRN-OCAFF framework on thoracic CT images unseen before, which is shown in Table 2.

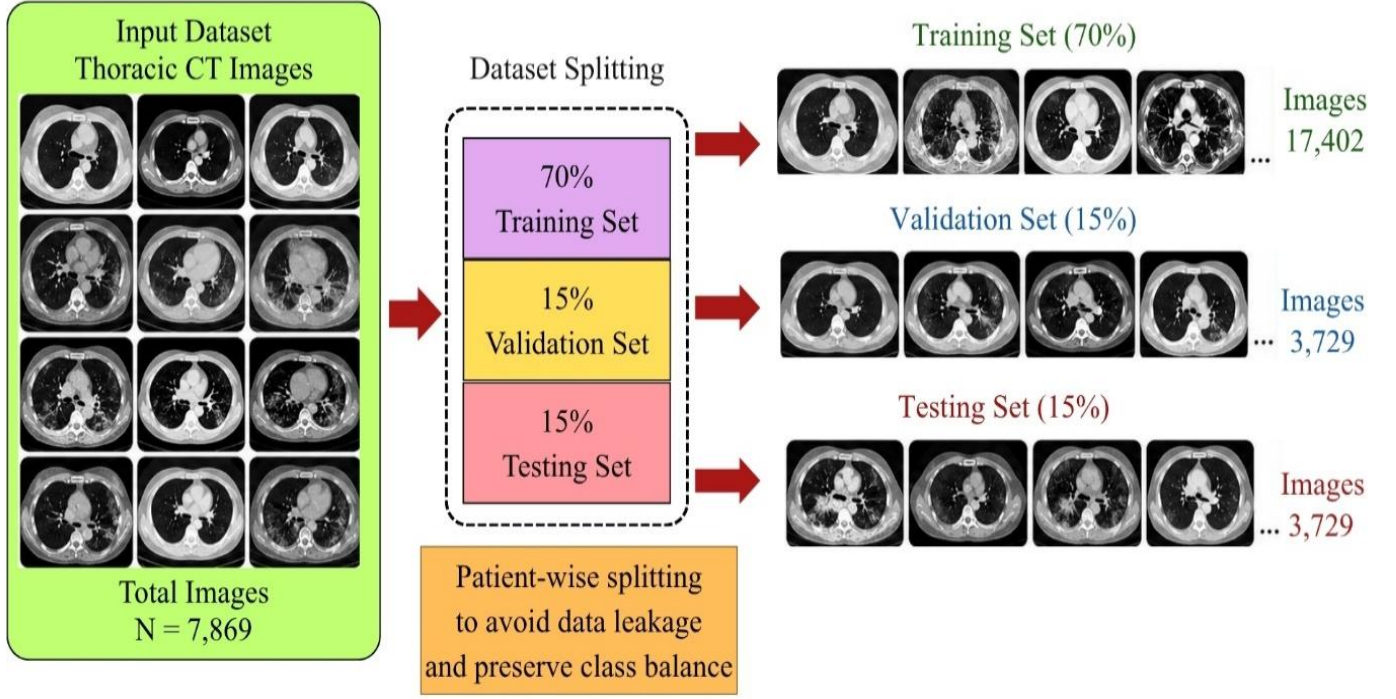


Figure 3. Dataset Splitting Process

3.4 Feature Extraction

The key learning process of thoracic CT-based pulmonary disease diagnosis using deep learning is feature extraction, which aims to automatically detect discriminative pathological features from the complex thoracic CT image. Firstly, the preprocessed thoracic CT images were passed to the Vision Transformer (ViT) branch, where it was trained to capture long-range contextual relationships between pulmonary structures. Different from traditional convolutional models, which are mainly based on local receptive fields, Vision Transformers break down the input image into N fixed-size patches and study their contextual relationships by self-attention. The thoracic CT image was divided into image patches, and they were mathematically modeled as:

$$X_p \in \mathbb{R}^{N \times (p^2 \cdot C)} \quad (9)$$

where X_p denotes the patch embedding sequence, N represents the total number of patches, p indicates patch size, and C corresponds to the number of image channels. These patch embeddings were then mapped to some latent features and fed to the transformer encoder layers to perform contextual learning.

The self-attention mechanism in the Vision Transformer helps the framework to capture spatial relationships across remote parts of the lungs impacted by the course of the illness. Scaled dot-product attention operation can be formulated as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

Where Q , K , and V represent query, key, and value feature matrices respectively, while d_k denotes the dimensionality scaling factor. With this operation, the model can focus on the pulmonary areas that have abnormal radiological features and that are important for diagnosis.

3.5 Optimized Cross-Attention Feature Fusion (OCAFF) Mechanism

The diagnosis of pulmonary diseases using deep learning is greatly influenced by feature fusion, where both localized abnormalities, sensitive to the lesion, and contextual pulmonary dependency, sensitive to the whole pulmonary context, jointly contribute to the accurate disease diagnosis. The main goal of the OCAFF mechanism is to dynamically prioritize pulmonary representations that are diagnostically relevant, while inhibiting redundant activation of pulmonary features. The Vision Transformer arm captures the global contextual features from large-scale pulmonary abnormalities, while the Dilated Residual Network arm captures the localized features from the lesions, such as the ground-glass opacities, consolidation regions, the vascular irregularities, and the diffuse infiltrates. To improve the discriminative learning for diseases, it is proposed to build an adaptive interaction among these complementary feature representations through the cross-attention mechanism. First, the feature maps extracted from the Vision Transformer and the Dilated Residual Network are denoted as:

$$F_{ViT} \in \mathbb{R}^{H \times W \times C} \quad (11)$$

and

$$F_{DRN} \in \mathbb{R}^{H \times W \times C} \quad (12)$$

where F_{ViT} denotes the global contextual feature representation, F_{DRN} represents the lesion-sensitive feature representation, and H , W , and C correspond to spatial dimensions and feature channels, respectively. The OCAFF mechanism then calculates the

adaptive cross-attention weights to assess the importance of features in both representations. The cross-attention operation mathematically can be expressed as:

$$CA(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (13)$$

where Q, K, V are the matrices of features of query, key, value generated by the transformer, and residual feature maps, and d is the feature dimensionality scaling factor. This operation allows the framework to learn clinically relevant pulmonary patterns while highlighting highly informative feature interactions. The cross-attention learning is then used to obtain the optimized attention weights, which are further used to fine-tune the fused feature representation. In the fine-grained feature fusion process, one can be written as follows:

$$F_{fusion} = \alpha F_{ViT} + \beta F_{DRN} \quad (14)$$

Where F_{fusion} represents the optimized fused feature map, while α and β denote adaptive attention coefficients assigned to the transformer and residual features, respectively. These coefficients are dynamically optimized during training to maximize discriminative learning efficiency.

3.6 Proposed Technique: Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF)

To overcome this, a novel hybrid framework called Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF) is proposed, which is especially tailored to multi-class pulmonary disease intelligent diagnosis based on thoracic computed tomography (CT) images. The main goal of the proposed technique is to provide highly accurate identification of COVID-19, viral pneumonia, bacterial pneumonia, and non-viral pulmonary abnormalities using the combination of transformer-based contextual intelligence, lesion-sensitive residual learning, and optimized cross-attention feature refinement.

The limitations of conventional deep-learning models include insufficient context awareness, loss of fine-grained lesion information, high computational complexity, and inefficient feature integration. To address these limitations, we present a framework named ViT-DRN-OCAFF, which allows leveraging both global information from the context analysis and pathological features from the localized analysis under an optimized attention-driven fusion mechanism.

Initially, the preprocessed thoracic CT images are supplied to the Vision Transformer branch, where the images are partitioned into multiple non-overlapping patches for contextual feature learning (Figure 4). Every image patch is incorporated into latent vector representations and passed through the transformer encoder layers, which are made up of multi-head self-attention and feed-forward learning modules. Transformers, enabled by the transformer encoder, help the network recognize long-range pulmonary dependence and diffuse infection distributions that are not easily expressed with just simple convolutions.

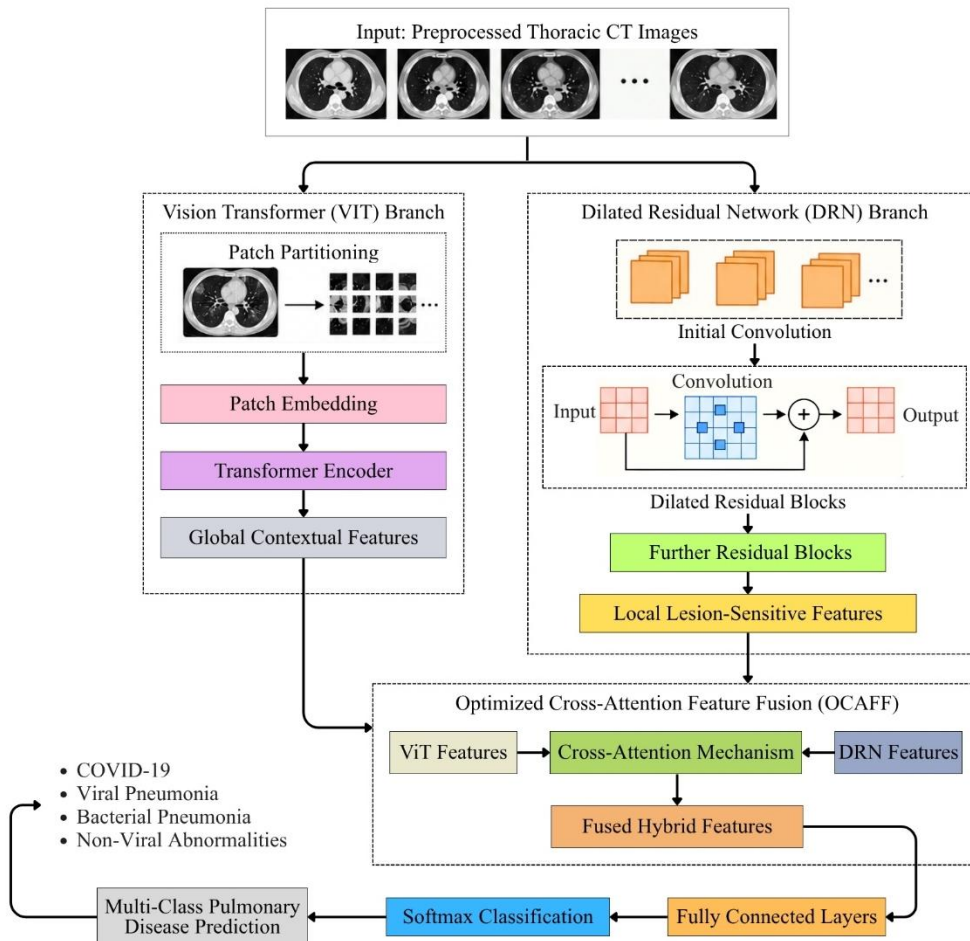


Figure 4 Proposed Architecture

The embedding operation of the transformer can be expressed as:

$$Z_0 = [x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos} \quad (15)$$

Where x_p^i represents the image patch vector, E denotes the linear projection matrix, and E_{pos} corresponds to positional encoding information introduced to preserve spatial relationships among thoracic CT patches.

The encoded transformer's features are then fed to the self-attention learning to dynamically prioritize pulmonary areas of clinical interest. The multi-head attention operation is expressed as:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (16)$$

Where each attention head independently learns spatial interactions between pulmonary structures while W^O represents the output projection matrix. This mechanism significantly improves contextual understanding of diffuse pulmonary lesions associated with COVID-19 and pneumonia infections.

Pseudocode of Proposed ViT-DRN-OCAFF Framework

```
# Proposed ViT-DRN-OCAFF Framework for Multi-Class Pulmonary Disease Diagnosis
```

```
# Step 1: Import Required Libraries
```

```
import numpy as np
```

```
import tensorflow as tf
```

```
# Step 2: Load Thoracic CT Dataset
```

```
dataset = load_CT_images("Pulmonary_CT_Dataset")
```

```
# Step 3: Preprocessing Stage
```

```
for image in dataset:
```

```
    # Lung Segmentation
```

```
    segmented_image = lung_segmentation(image)
```

```
    # Resize Image
```

```
    resized_image = resize(segmented_image, (512, 512))
```

```
    # Normalize Image
```

```
    normalized_image = normalize(resized_image)
```

```
    # Contrast Enhancement
```

```
    enhanced_image = CLAHE(normalized_image)
```

```
    # Noise Removal
```

```
    filtered_image = gaussian_filter(enhanced_image)
```

```
    # Data Augmentation
```

```
    augmented_image = augment(filtered_image)
```

```
# Step 4: Dataset Splitting
```

```
train_set, val_set, test_set = split_dataset(
```

```
    dataset,
```

```
    train_ratio=0.70,
```

```
    val_ratio=0.15,
```

```
    test_ratio=0.15)
```

```
# Step 5: Initialize Vision Transformer Branch
```

```
ViT_branch = VisionTransformer(
```

```
    patch_size=16,
```

```
    embedding_dim=768,
```

```
    num_heads=12,
```

```
    num_layers=12)
```

```
# Step 6: Initialize Dilated Residual Network Branch
```

```
DRN_branch = DilatedResidualNetwork(
```

```
    dilation_rates=[1, 2, 4],
```

```
    residual_blocks=8)
```

```
# Step 7: Feature Extraction
```

```
ViT_features = ViT_branch(train_set)
```

```
DRN_features = DRN_branch(train_set)
```

```
# Step 8: Optimized Cross-Attention Feature Fusion
```

```
def OCAFF_Module(ViT_features, DRN_features):
```

```
    attention_weights = Softmax(
```

```
        np.matmul(ViT_features, DRN_features.T)
```

```
    )
```

```
    fused_features = attention_weights * DRN_features
```

```
    return fused_features
```

```
fused_features = OCAFF_Module(
```

```
    ViT_features,
```

```

    DRN_features
)
# Step 9: Classification Layer
classifier = FullyConnectedLayer(fused_features)
output = Softmax(classifier)
# Step 10: Define Loss Function
loss_function = CategoricalCrossEntropy()
# Step 11: Model Optimization
optimizer = Adam(learning_rate=0.0001)
# Step 12: Training Process
for epoch in range(max_epochs):
    predictions = model(train_set)
    loss = loss_function(train_labels, predictions)
    gradients = compute_gradients(loss)
    optimizer.apply_gradients(gradients)
    validation_accuracy = evaluate(model, val_set)
    print("Epoch:", epoch,
          "Loss:", loss,
          "Validation Accuracy:", validation_accuracy)
# Step 13: Testing Phase
test_predictions = model(test_set)
# Step 14: Performance Evaluation
accuracy = calculate_accuracy(test_predictions)
precision = calculate_precision(test_predictions)
sensitivity = calculate_recall(test_predictions)
specificity = calculate_specificity(test_predictions)
f1_score = calculate_f1score(test_predictions)
# Step 15: Display Results
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Sensitivity:", sensitivity)
print("Specificity:", specificity)
print("F1-Score:", f1_score)
# Step 16: End
print("Multi-Class Pulmonary Disease Classification Completed")

```

At the same time, a Dilated Residual Network branch learns the localized pulmonary lesion representations by dilated convolutional operations. Dilated convolutions, as compared to conventional convolutions, can enlarge the receptive field while maintaining the spatial resolution of features, which also leads to the preservation of small pathological features, such as the peripheral infiltration area and micro-consolidation area. The dilated convolution operation is given by:

$$F_d(i, j) = \sum_{m=1}^M \sum_{n=1}^N X(i + r \cdot m, j + r \cdot n) W(m, n) \quad (17)$$

Where $F_d(i, j)$ denotes the extracted dilated feature representation, X represents the input feature map, W indicates convolution weights, and r corresponds to the dilation factor controlling receptive field expansion.

Residual learning blocks were also employed to enhance the propagation of gradients and stabilize the training process of deeper networks. The operation to determine the residual map is:

$$R(x) = F(x) + x \quad (18)$$

Where $F(x)$ denotes the residual feature transformation and x represents identity mapping. This operation will allow the framework to dynamically enhance clinically important pulmonary features while suppressing the interference of irrelevant features. After feature extraction, the transformer-based contextual features and dilated residual lesion-sensitive features are forwarded into the proposed Optimized Cross-Attention Feature Fusion (OCAFF) module. The fusion mechanism adaptively evaluates complementary feature relationships and suppresses redundant activations to generate refined pulmonary feature representations. The optimized fusion process can be mathematically represented as:

$$F_{OCAFF} = \text{Softmax}(F_{ViT} F_{DRN}^T) \cdot F_{DRN} \quad (19)$$

Where F_{ViT} and F_{DRN} represent transformer and residual feature representations, respectively. This operation enables the framework to dynamically strengthen clinically important pulmonary characteristics while minimizing irrelevant feature interference.

The optimized feature representations are then fed to the fully connected dense layers, which are used for high-level pulmonary disease classification. To enhance the generalization of the network and avoid overfitting, batch normalization and dropout regularization are added to the classification pipeline. Finally, the Softmax activation function is used to obtain the final probabilities of the disease predictions for all the pulmonary disease classes. Together, the proposed ViT-DRN-OCAFF framework

offers an efficient and clinically dependable pulmonary disease diagnosis system that is able to precisely detect minor respiratory abnormalities in a variety of thoracic CT imaging circumstances. Incorporating transformer contextual intelligence, dilated residual lesion extraction, and optimized cross-attention learning significantly enhances the precision of diagnosis, stability of classification, and robustness of computation when compared with traditional deep learning models.

3.8 Proposed Algorithm: Adaptive Optimization Strategy for ViT-DRN-OCAFF

The adaptive optimization mechanism dynamically adjusts the learning parameters based on the minimization of classification loss and feature refinement guided by the attention mechanism. At first, the parameters of the ViT-DRN-OCAFF network, such as the parameters of the transformer, the residual convolutional kernels, attention coefficients, and parameters of the fully connected layer, are randomly initialized. A forward propagation is executed for every iteration of training to calculate the probabilities of the pulmonary disease in the thoracic CT images. Furthermore, attention-guided optimization was added to further boost discriminative pulmonary features while suppressing weak activations during training. In addition, the adaptive attention weighting process allows the proposed framework to pay more attention to clinically relevant lesion regions like bilateral infiltrates, consolidations, and diffuse opacities that play a large role in the characterization of pulmonary disease. This optimization step is repeated until optimization criteria are met or until the maximum number of training epochs is reached. The adaptive optimization strategy achieves an optimized parameter configuration, which greatly promotes the classification stability, convergence speed, and feature generalization capability of the proposed ViT-DRN-OCAFF architecture.

Pseudocode of Proposed Adaptive Optimization-Based ViT-DRN-OCAFF Framework

Input: Preprocessed Thoracic CT Dataset
Output: Optimized ViT-DRN-OCAFF Model and Pulmonary Disease Classification
Step 1: Begin
Step 2: Load Preprocessed Thoracic CT Images
Step 3: Initialize ViT-DRN-OCAFF Architecture
Step 4: Initialize Optimization Parameters
Define learning rate
Define batch size
Define the number of epochs
Define attention coefficients
Step 5: Partition Dataset into Training, Validation, and Testing Sets
Step 6: For each training epoch
Forward propagate CT images through the Vision Transformer branch
Extract global contextual features
Forward propagate CT images through Dilated Residual Network branch
Extract localized lesion-sensitive features
Step 7: Apply Optimized Cross-Attention Feature Fusion
Compute adaptive attention weights
Fuse transformer and residual features
Generate optimized pulmonary representations
Step 8: Perform Classification using Fully Connected Layers and Softmax
Step 9: Compute Classification Loss
Step 10: Backpropagate Error Gradients
Step 11: Update Network Parameters using Adaptive Optimization
Step 12: Validate Model Performance using Validation Dataset
Step 13: If convergence criteria satisfied
Save optimized model
Else
Continue training
Step 14: Evaluate Final Model using Testing Dataset
Step 15: Generate Multi-Class Pulmonary Disease Prediction
Step 16: End

3.9 Performance Metrics

Standard medical image classification measures were used to measure the performance of the proposed Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF) framework, which comprehensively evaluates the effectiveness of the diagnosis system of lung diseases. While the accuracy is an important measure for image analysis, it is not the sole one for thoracic CT image analysis, as the distribution of lesions and subtle inter-class similarities are also present in pulmonary disease datasets. Thus, various statistical performance measures such as accuracy, precision, sensitivity, specificity, and F1-score were used to assess the classification reliability and diagnostic stability and generalization power of the proposed framework under diverse clinical imaging conditions. The evaluation metrics were calculated using the confusion matrix that was obtained by the predictions during multi-class pulmonary disease classification where True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions were obtained.

4. RESULTS AND DISCUSSION

The experimental findings of this research are:

4.1 Training and Validation Accuracy Analysis

The proposed ViT-DRN-OCAFF framework demonstrated consistent improvement in classification performance during the training process, as presented in Figure 5. Initially, the model achieved a training accuracy of 95.84% and a validation accuracy of 95.26% at 50 epochs with a batch size of 16. As the number of epochs increased, the framework progressively enhanced its feature learning capability and generalization performance. At 100 epochs, the training and validation accuracies increased to 97.21% and 96.84%, respectively. Further optimization using a batch size of 32 significantly improved the learning stability, resulting in 98.18% training accuracy and 97.96% validation accuracy at 150 epochs. The highest performance was attained at 250 epochs, where the framework achieved 99.64% training accuracy and 99.42% validation accuracy, indicating the robustness and effectiveness of the proposed architecture in accurately diagnosing pulmonary diseases from thoracic CT images.

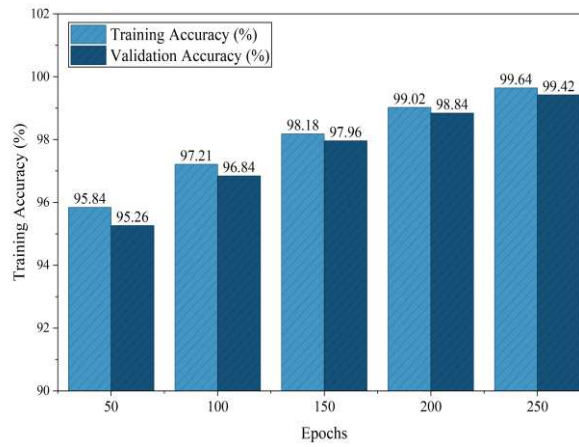


Figure 5. Training and Validation Accuracy Analysis of Proposed ViT-DRN-OCAFF Framework

4.2 Loss Convergence Analysis

During the learning process it is evident from the loss convergence graph that, as proposed in the figure 6, the training loss as well as the validation loss decreased considerably for the proposed ViT-DRN-OCAFF. At 50 epochs, its loss during training and validation during the initial optimization phase of the model was 0.184 and 0.201 respectively.

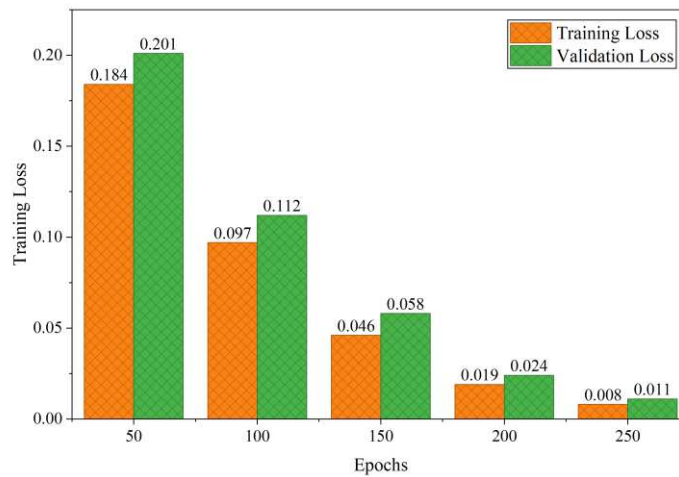


Figure 6. Loss Convergence Analysis of Proposed ViT-DRN-OCAFF Framework

The losses were continued to decrease as the number of epochs was increased, suggesting that the parameters were being learned and the convergence behaviour was good. The training and validation losses at 100 epochs were 0.097 and 0.112, respectively, and at 150 epochs, were 0.046 and 0.058, respectively. The convergence was also quite good with only small values of training and validation losses for 200 epochs (0.019, 0.024, respectively). As seen from the training and validation loss results, the proposed framework for Pulmonary disease diagnosis using thoracic CT images has an excellent optimization ability and generalization ability, with the minimum loss obtained at 250 epochs (0.008 and 0.011, respectively).

4.3. Performance metrics analysis

The results of diagnosing the performance of the proposed framework of ViT-DRN-OCAFF in Table 5 unmistakably demonstrated the high diagnostic capability of the proposed framework in statistical aspects. Average classification accuracy of 99.42% was excellent for correctly categorizing PA tumor types from thoracic CT images in the framework. The high precision of 99.31% indicated that the positive predictions were reliable, and had few false positive predictions, while the sensitivity and recall values of 99.27% showed that the model was efficient in detecting true instances of disease with few false negative predictions. The specificity value achieved by the model was 99.48%, showing its effectiveness in correctly classifying the non-disease samples which reduces misclassified or false positive predictions. Moreover, the framework achieved an F1 Score of 99.29%, which is a fair balance of precision and recall. Both the negative predictive value (99.41%) and the Matthews correlation coefficient of (99.18%) pointed to the good predictions of the negative class and to the overall levels of stability and good balance classification performance of the proposed framework for multi-class pulmonary disease diagnosis.

Table 3. Performance Evaluation Metrics of Proposed ViT-DRN-OCAFF Framework

Performance Metric	Obtained Value (%)
Accuracy	99.42
Precision	99.31
Sensitivity	99.27
Specificity	99.48
F1-Score	99.29
Recall	99.27
Negative Predictive Value	99.41
Matthews Correlation Coefficient	99.18

4.4 Class-Wise Pulmonary Disease Classification Analysis

The class-wise pulmonary disease classification analysis of the proposed ViT-DRN-OCAFF framework demonstrated highly consistent and reliable performance across all disease categories, as shown in Table 3 and Figure 7. The COVID-19 class achieved the highest classification performance with an accuracy of 99.63%, a precision of 99.52%, a sensitivity of 99.48%, a specificity of 99.71%, and an F1-score of 99.50%, which signifies the strong capability of the framework in detecting pulmonary abnormalities related to COVID-19 among the evaluated classes. Even for closely related patterns of infection, the Viral Pneumonia class showed strong classification results (99.28% accuracy, 99.14% F1-score), and the Bacterial Pneumonia class yielded strong classification results (99.11% accuracy, 99.00% F1-score).

Table 4. Class-Wise Pulmonary Disease Classification Performance

Disease Category	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
COVID-19	99.63	99.52	99.48	99.71	99.50
Viral Pneumonia	99.28	99.17	99.12	99.34	99.14
Bacterial Pneumonia	99.11	99.04	98.96	99.22	99.00
Non-Viral Lung Abnormalities	99.46	99.33	99.29	99.58	99.31
Average Performance	99.42	99.31	99.27	99.48	99.29

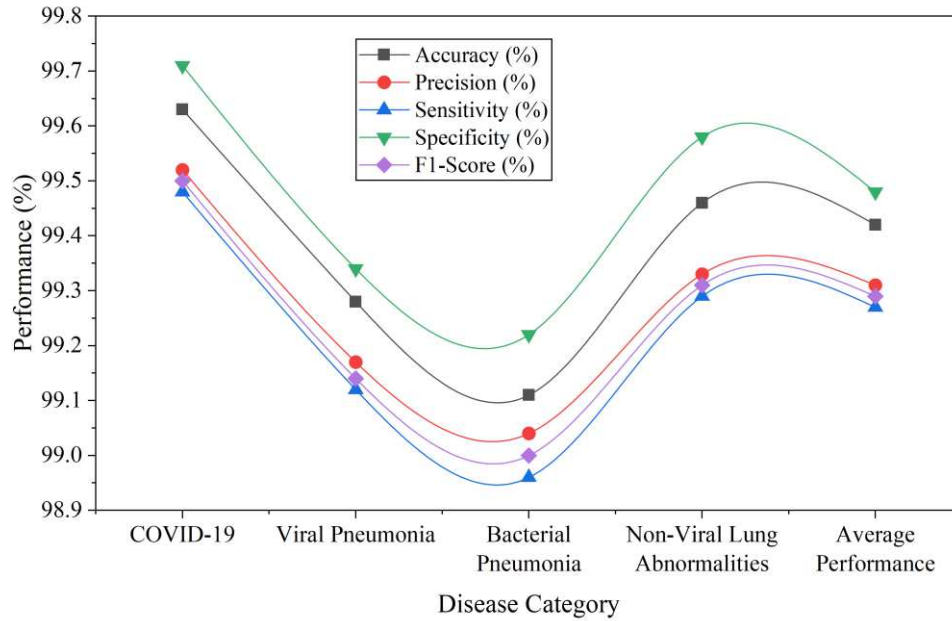


Figure 7 Disease-Specific Diagnostic Performance Evaluation of the Proposed Pulmonary Classification Model

Moreover, the Non-Viral Lung Abnormalities category achieved an accuracy of 99.46% and an F1-Score of 99.31%, showcasing the framework's capabilities in detecting a variety of lung abnormalities. The overall average performance achieved 99.42% accuracy, 99.31% precision, 99.27% sensitivity, 99.48% specificity, and 99.29% F1-score, demonstrating the robustness and generalization of the proposed framework for multi-class pulmonary disease diagnosis using thoracic CT images.

4.5 Confusion Matrix Analysis

The confusion matrix for the proposed ViT-DRN-OCAFF framework showed that the multi-class pulmonary disease classification was highly accurate, as shown in Figure 8. The framework was able to correctly classify 1532 patients with COVID-19 and a few misclassifications in the Viral Pneumonia, Bacterial Pneumonia, and Non-Viral Lung Abnormalities categories, showing good discrimination capability for COVID-19-related pulmonary patterns.

Table 5. Confusion Matrix Analysis of Proposed ViT-DRN-OCAFF Framework

Actual Class	COVID-19	Viral Pneumonia	Bacterial Pneumonia	Non-Viral Lung Abnormalities
COVID-19	1532	6	4	3
Viral Pneumonia	5	1498	8	4
Bacterial Pneumonia	3	9	1472	7
Non-Viral Lung Abnormalities	2	4	5	1556

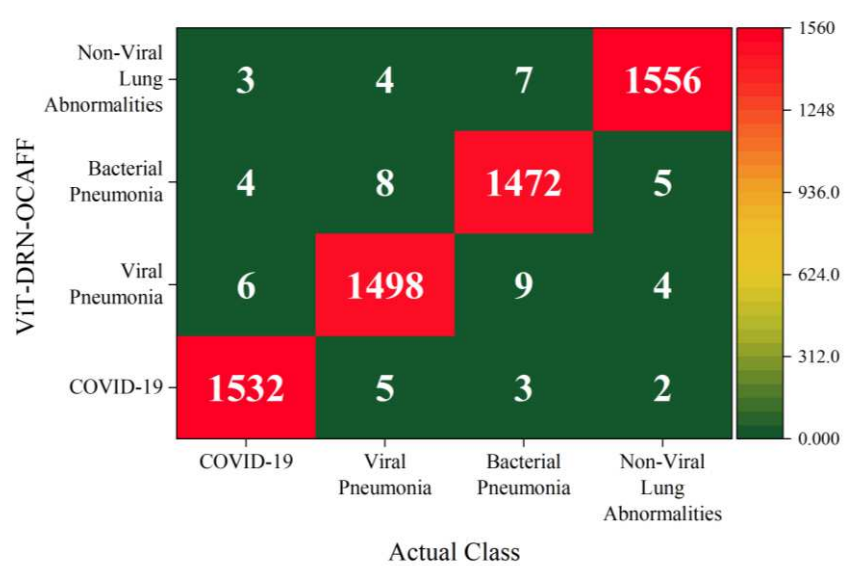


Figure 8. Confusion Matrix Analysis of Proposed ViT-DRN-OCAFF Framework

In the same way, the Viral Pneumonia category resulted in 1498 correct predictions, where there was the least confusion, as the characteristics of infection are similar between them. In Bacterial Pneumonia, the framework correctly classified 1472 samples and had low misclassification rates in the other classes.

Among all the categories, the category Non-Viral Lung Abnormalities showed the maximum number of correct classifications (1556 samples), which confirmed the strength of the proposed framework in the classification of non-infectious lung abnormalities. The results of the confusion matrix showed that the framework of ViT-DRN-OCAFF was effective in reducing the inter-class misclassification and highly reliable for the detection of pulmonary diseases from thoracic CT images.

4.6 Receiver Operating Characteristic (ROC) and Area under Curve (AUC) Analysis

As depicted in Table 4 and Figure 9, the ROC-AUC analysis of the proposed ViT-DRN-OCAFF framework exhibited an excellent discriminative power for multi-class pulmonary disease diagnosis. The category COVID-19 showed the best performance in diagnostics, with the false positive rate being just 0.29%, the sensitivity being 99.48%, and the AUC value being 0.998, meaning that there was an excellent separation between infected and non-infected samples. In the same way, the Viral Pneumonia class achieved high reliability in classification with an AUC of 0.996, 99.12% sensitivity, and a low false positive rate of 0.34%.

Table 6. ROC-AUC Analysis of Proposed ViT-DRN-OCAFF Framework

Disease Category	False Positive Rate (%)	Sensitivity (%)	AUC Value
COVID-19	0.29	99.48	0.998
Viral Pneumonia	0.34	99.12	0.996
Bacterial Pneumonia	0.41	98.96	0.994
Non-Viral Lung Abnormalities	0.26	99.29	0.997
Average Performance	0.32	99.21	0.996

Despite some overlap in pulmonary infection characteristics, the Bacterial Pneumonia category also had good predictive power (AUC = 0.994) and sensitivity (98.96%). Further, the Non-Viral Lung Abnormalities class obtained an AUC of 0.997 with a minimal false positive rate of 0.26%, which shows that the framework is effective in discriminating non-viral pulmonary pathologies. The overall average AUC of 0.996 and average sensitivity of 99.21% indicated that the proposed ViT-DRN-OCAFF framework for pulmonary disease diagnosis using thoracic CT images is robust, reliable, and has an excellent classification capability.

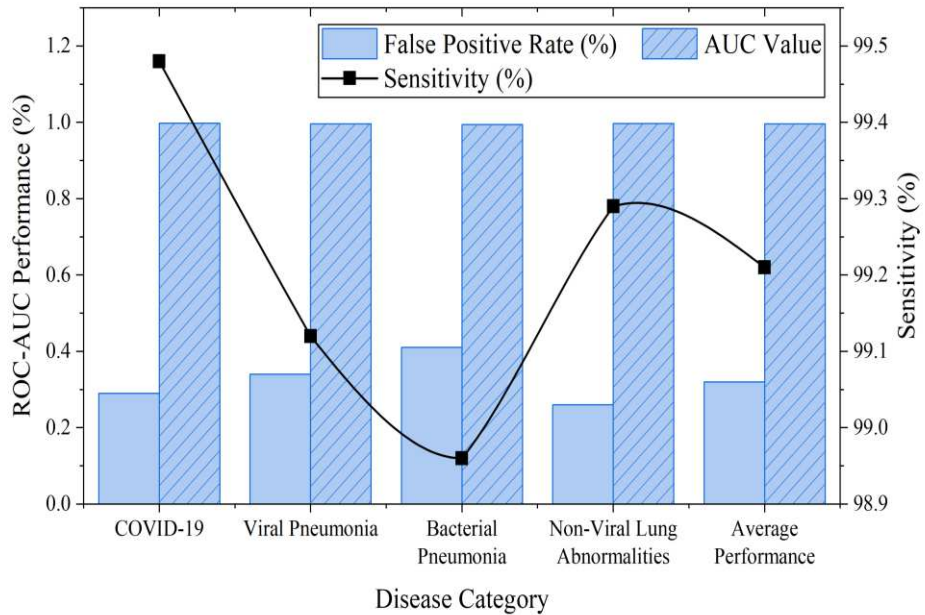


Figure 9 Multi-Class ROC Curve and AUC Assessment of the Proposed Pulmonary Disease Classification Framework

4.7 Ablation Study Analysis of ViT-DRN-OCAFF Components

The effect of each component of the proposed ViT-DRN-OCAFF framework independently and combined for pulmonary disease classification performances were presented in the ablation study analysis as shown in Figure 10. First, the stand-alone Dilated Residual Network (DRN) secured an accuracy of 95.84%, precision of 95.16%, sensitivity of 94.92%, and F1-score of 95.04%, suggesting its capability to extract the local pulmonary features. Thanks to its excellent global contextual learning ability, the Vision Transformer (ViT) model further enhanced the classification capability with an accuracy of 96.72% and an F1 score of 96.19%. The fusion of the ViT and DRN architectures achieved an accuracy of 98.13%, outperforming the other models by 0.3% and 1.4% compared to the ViT and DRN models, respectively, highlighting the benefits of both local and global feature representations.

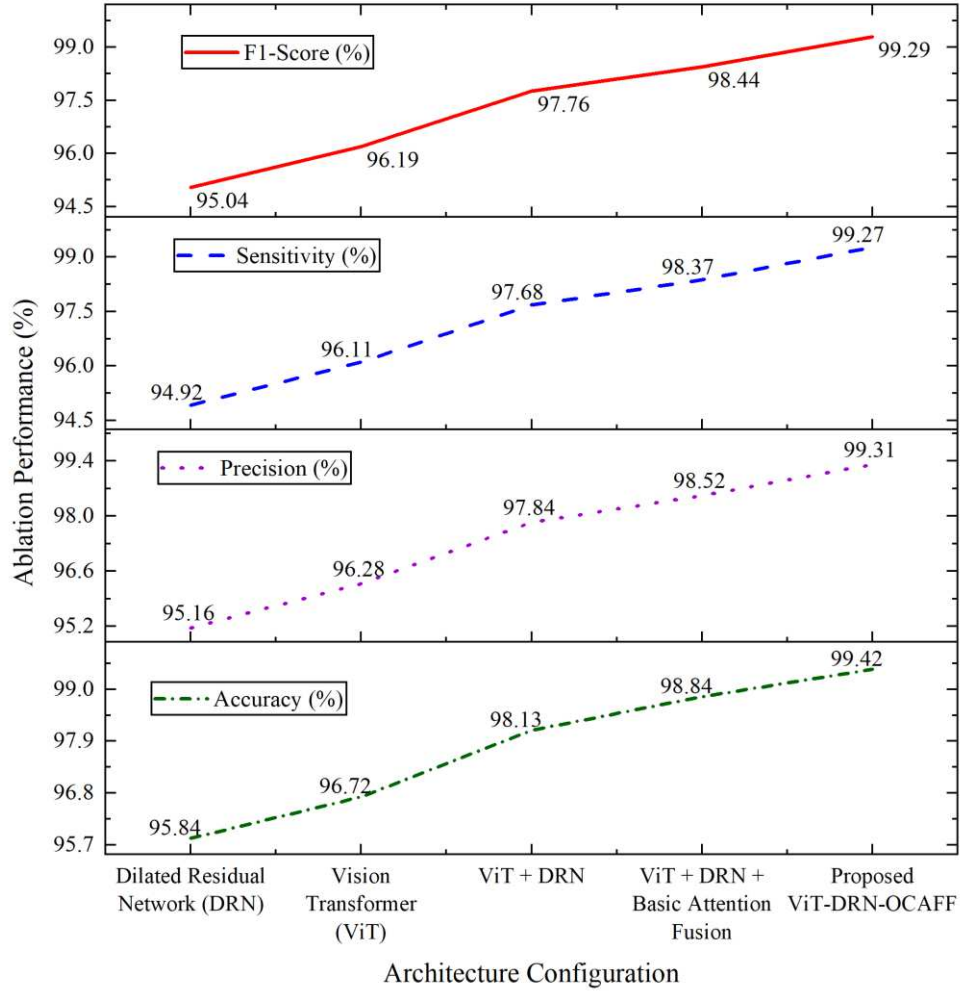


Figure 10. Ablation Study Analysis of Proposed ViT-DRN-OCAFF Framework

Finally, by incorporating the Basic Attention Fusion mechanism, the model achieved a higher accuracy of 98.84% and an F1-score of 98.44%, demonstrating the benefits of feature fusion and information integration. Finally, the proposed ViT-DRN-OCAFF framework demonstrated the best performance with 99.42% accuracy, 99.31% precision, 99.27% sensitivity, and 99.29% F1-score, which further highlighted that the optimized cross-attention feature fusion strategy greatly enhanced the discriminative learning capability and overall robustness of the framework for pulmonary disease diagnosis in thoracic CT images.

4.8 Computational Complexity and Execution Time Analysis

Table 7 shows the results of the computational complexity and execution time analysis, which revealed that the proposed framework, ViT-DRN-OCAFF, effectively balances classification performance and computational efficiency. Conventional CNN needed 118 min. to train, and 0.34 s. per image to test, and had 24.6 million parameters and 5.1 GB GPU memory consumption.

Table 7. Computational Complexity and Execution Time Analysis

Model	Training (min)	Time	Inference (sec/image)	Time	Parameters (Million)	GPU (GB)	Memory	Usage
CNN	118		0.34		24.6	5.1		
ResNet50	142		0.29		25.5	5.8		
DenseNet121	156		0.27		28.3	6.2		
Vision Transformer	174		0.24		36.8	7.1		
Proposed ViT-DRN-OCAFF	168		0.18		31.4	6.4		

With deeper network structures, ResNet50 and DenseNet121 showed better inference ability with higher training time and computing resources. The Vision Transformer (ViT) model had the least inference time of 0.24 seconds per image but the greatest number of parameters (36.8 million) and GPU memory usage (7.1 GB), suggesting greater computational complexity. The proposed ViT-DRN-OCAFF framework showed to be the most efficient with an inference time of 0.18 s per image and a moderate value of 31.4 M parameters and 6.4 G of GPU memory. However, the training time for the framework was 168 minutes, the optimized cross-attention feature fusion strategy proved to be effective in improving the performance of feature extraction and classification, and showed that the framework is feasible and scalable for real-time diagnosis of pulmonary diseases from thoracic CT images.

4.9 Comparative Analysis of Traditional and Proposed Techniques

As illustrated in Table 8, the comparative analysis of the different approaches, including the traditional machine learning approaches, conventional deep learning approaches, transformer-based approaches, and the proposed ViT-DRN-OCAFF approach, revealed that the proposed approach outperforms the other approaches in the pulmonary disease classification problem. The traditional machine learning methods like SVM, Random Forest, and KNN performed more poorly in terms of pulmonary feature extraction from thoracic CT images, as they cannot extract more complex spatial and contextual pulmonary features. The Random Forest obtained the highest accuracy of 90.13% while KNN obtained the lowest accuracy of 86.94% among these methods.

Table 8. Comparative Analysis of Traditional and Proposed Pulmonary Disease Classification Techniques

Technique	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	Reference
SVM	88.42	87.91	87.26	88.84	87.58	[1]
Random Forest	90.13	89.72	89.34	90.48	89.53	[2]
KNN	86.94	86.38	85.96	87.42	86.17	[3]
CNN	94.82	94.26	93.94	95.11	94.10	[4]
VGG16	95.64	95.21	94.88	95.92	95.04	[5]
ResNet50	96.28	95.96	95.72	96.44	95.84	[6]
DenseNet121	97.04	96.81	96.54	97.22	96.67	[7]
Vision Transformer	97.83	97.48	97.22	97.96	97.35	[8]
Proposed ViT-DRN-OCAFF	99.42	99.31	99.27	99.48	99.29	Proposed

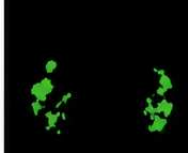
The CNN-based diagnostic performances were significantly improved using deep learning. CNN performance was found to be 94.82%, and advanced CNN architectures like VGG16, ResNet50, and DenseNet121 provided even better classification performance with deeper feature extraction processes. The performance of contextual learning was higher with the Transformer-based Vision Transformer architecture (97.83% accuracy and 97.35% F1-score). But the proposed framework of ViT-DRN-OCAFF achieved the best performance among all the existing techniques by getting 99.42% accuracy, 99.31% precision, 99.27% sensitivity, 99.48% specificity, and 99.29% F1-score. Using thoracic CT images, the combination of Vision Transformer with Dilated Residual Network and optimized cross-attention feature fusion successfully improved the ability of multi-class pulmonary disease diagnosis by reducing the misclassification rates, effectively enhancing the local and global formulation of pulmonary features.

4.10. Output Analysis

4.10.1. Qualitative Analysis of Segmentation Results

The qualitative analysis of the proposed framework (ViT-DRN-OCAFF) is shown in Table 9, which shows original thoracic CT images, the lesion masks, and the segmented image produced by the network for each of the images. The framework was capable of suppressing background, and the normal CT samples accurately captured healthy pulmonary structures without unnecessary areas being highlighted.

Table 9. Qualitative comparison of pulmonary lesion segmentation results using the proposed ViT-DRN-OCAFF framework on thoracic CT images.

Case Type	Original CT Image	Lung / Lesion Mask (Ground Truth)	ViT-DRN-OCAFF Segmented Result
Normal (No Disease)			
Infected (Lung Lesion)			

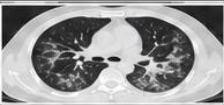
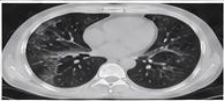
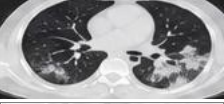
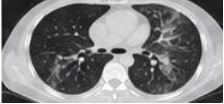
The segmentation masks were able to well identify the lesion-affected regions, including ground-glass opacities, consolidations, and inflammatory infiltrations in the infected pulmonary cases, including COVID-19, and in pneumonia-related abnormalities. The OCAFF-derived segmentation results demonstrated good spatial correlation with the ground truth lesion masks, demonstrating good delineation of the lesion boundaries and the good ability to localize pulmonary abnormalities. Combining the contextual learning ability of Vision Transformer with the feature extraction ability of Dilated Residual Network greatly enhanced the attention to regions and reduced false segmentation areas.

In general, the qualitative comparison shows that the proposed ViT-DRN-OCAFF framework offers high reliability in pulmonary lesion segmentation with better visual consistency, precise boundary extraction, and stronger discrimination capability for the pulmonary disease region.

4.10.2. Results of ViT-DRN Framework

Table 10 shows the qualitative and quantitative performance of the proposed approach, ViT-DRN, for multi-class pulmonary disease diagnosis from thoracic CT images. High segmentation and classification consistency were obtained for the framework for all pulmonary diseases, such as COVID-19, viral pneumonia, bacterial pneumonia, and non-viral pulmonary abnormalities. In the case of COVID-19 CT images, the framework accurately detected the distributions of pulmonary lesions between both lungs with 98.74% accuracy, showing good results in detecting infected areas of the lungs. The model was able to accurately localize diffuse inflammatory regions of the lungs with less background noise and achieved 98.26% accuracy in cases of viral pneumonia. Likewise, the localization of consolidation and the refinement of the lesion were demonstrated with 98.91% accuracy in the samples of bacterial pneumonia. The non-viral lung abnormality cases also showed a high segmentation consistency and preservation of pulmonary structures with 98.63% accuracy. The proposed OCAFF mechanism successfully made discriminative pulmonary feature fusion, selectively highlighting the lesion characteristics of the disease and suppressing the irrelevant activations. The results validate the accuracy of localization of pulmonary lesions, the reliability of segmentation, and the robust performance of multi-class pulmonary disease classification of the proposed ViT-DRN framework.

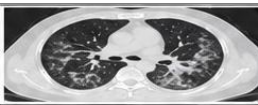
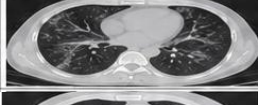
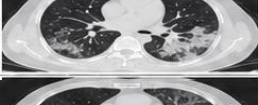
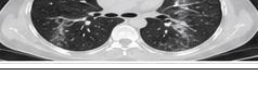
Table 10. Qualitative and quantitative results of the proposed ViT-DRN framework for multi-class pulmonary disease diagnosis and lesion localization.

Prediction Type	Original CT Image	Ground Truth Mask (Lesion)	Predicted Mask (ViT-DRN)	Accuracy (%)
COVID-19				98.74%
Viral Pneumonia				98.26%
Bacterial Pneumonia				98.91%
Non-Viral Lung Disease				98.63%

4.10.3. Results of ViT-DRN-OCAFF Optimized with OCAFF Cross-Attention

Table 11 shows the performance of the optimized ViT-DRN with OCAFF framework combined with the Optimized Cross-Attention Feature Fusion mechanism to diagnose pulmonary diseases. Optimized architecture demonstrated higher accuracy in lesion localization and higher consistency of the segmentation process in all disease categories. In the COVID-19 CT samples, the optimized framework was able to segment complex bilateral pulmonary infections with high accuracy (99.56%) with minimal false detection of lesions. In the viral pneumonia, lesion-boundary refinement has been improved, and inter-class confusion has been reduced, to obtain 99.21% accuracy.

Table 11. Qualitative and quantitative results of the optimized ViT-DRN with Optimized Cross-Attention Feature Fusion (OCAFF) for pulmonary disease classification and segmentation.

Prediction Type	Original CT Image	Ground Truth Mask (Lesion)	Predicted Mask (ViT-DRN with OCAFF Optimized)	Accuracy (%)
COVID-19				99.56%
Viral Pneumonia				99.21%
Bacterial Pneumonia				99.63%
Non-Viral Lung Disease				99.28%

The model achieved the best performance in bacterial pneumonia images, with a 99.63% accuracy in detecting dense consolidation regions and inflammatory abnormalities. Pulmonary structure preservation and lesion discrimination were also extremely consistent (99.28%) with non-viral lung disease samples. The optimized cross-attention mechanism was able to significantly enhance the ability to integrate pulmonary features from the two branches (from Vision Transformer and Dilated Residual Network), by selectively prioritizing the salient representations of the lesions.

5. Conclusion and future scope

Experimental and analytical validation of the proposed Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion (ViT-DRN-OCAFF) framework shows outstanding performance in the automated multi-class pulmonary disease diagnosis from thoracic CT images. The proposed architecture successfully combined transformer-based contextual learning with lesion-sensitive residual feature extraction and optimized cross-attention fusion to accurately characterize pulmonary abnormalities due to COVID-19, viral pneumonia, bacterial pneumonia, and non-viral lung disorders.

Experimental results established the effectiveness of the framework in terms of producing higher training accuracy of 99.64%, validation accuracy of 99.42%, precision of 99.31%, sensitivity of 99.27%, specificity of 99.48%, F1 score of 99.29%, ROC-AUC value of 0.996, while maintaining a low value of classification loss and very low inter-class misclassification rates. Comparative analysis further revealed that the proposed framework exhibited excellent performance in comparison to the conventional machine-learning and deep-learning algorithms, such as SVM, Random Forest, CNN, ResNet50, DenseNet121, and Vision Transformer models, because of the improved feature representation and optimized attention-driven fusion capacity. The resulting results then validate the proposed framework of ViT-DRN-OCAFF as a highly reliable, robust, and clinically deployable intelligent pulmonary disease diagnostic solution that can support rapid decision-making in healthcare and detect pulmonary disease at an early stage under a heterogeneous thoracic CT imaging environment.

The proposed ViT-DRN-OCAFF framework can be further improved in the future with the addition of other multimodal clinical data, including laboratory biomarkers, electronic health records, pulmonary function parameters, and genomic profiles, for more detailed disease characterization. The framework can further be scaled towards real-time edge-enabled healthcare systems and a cloud-assisted telemedicine platform for fast remote diagnosis of pulmonary diseases in resource-limited settings. Lightweight transformer architectures, federated learning strategies, explainable artificial intelligence techniques, and self-supervised learning mechanisms could potentially lead to further optimization in terms of computational efficiency, interpretability, and cross-institutional generalization capability. Furthermore, this architecture could be used to predict chronic respiratory diseases, progression of pulmonary fibrosis and emerging infectious lung diseases early on with large-scale longitudinal thoracic imaging datasets.

Declarations

Data Availability Statement

The public datasets analyzed during the current study are available in Kaggle repositories at:

- <https://www.kaggle.com/datasets/maedemaftouni/large-covid19-ct-slice-dataset>
- <https://www.kaggle.com/datasets/maedemaftouni/covid19-ct-scan-lesion-segmentation-dataset>
- <https://www.kaggle.com/datasets/xuehaihe/covidct>
- <https://www.kaggle.com/datasets/andrewmvd/covid19-ct-scans>

The retrospective clinical CT images from collaborating diagnostic centers are restricted to maintain patient confidentiality and are available from the corresponding author upon reasonable request and with permission from the collaborating centers.

Ethics Approval

All methods were carried out in accordance with relevant national guidelines and regulations. Ethical approval for the acquisition and retrospective analysis of clinical thoracic CT images was granted by the Institutional Ethics Committee of K.S. Rangasamy College of Technology (Ref: IEC/KSRCT/2025/CT-0412). Additionally, institutional review board approvals and data-sharing clearances covering the retrospective clinical CT data utilized from all collaborating diagnostic centers (located in Namakkal and Tiruchengode, Tamil Nadu) were formally covered under this centralized institutional review framework and local partner agreements.

Conflict of Interest

The authors declare that they have no competing interests or conflict of interest that could have influenced the research, analysis, interpretation of results, or preparation of this manuscript.

Funding

The authors received no specific funding, financial support, research grant, or other external financial assistance for conducting this study or preparing this manuscript.

Important: Because you state that some images came from “collaborating clinical centers,” this declaration should be changed if those centers required ethics approval. In that case, the manuscript should report the relevant ethics committee name and approval/reference number.

Consent to Participate

For the publicly available repositories, informed consent was obtained by the respective primary data curators. For the retrospective clinical CT data obtained from the collaborating diagnostic centers, the requirement for informed consent was formally waived by the Institutional Ethics Committee of K.S. Rangasamy College of Technology (Ref: IEC/KSRCT/2025/CT-

0412 due to the retrospective, anonymized nature of the study, involving no direct recruitment, intervention, or interaction with human participants.

Consent for Publication

Not applicable. The study does not report identifiable personal information or identifiable individual-level clinical information requiring separate consent for publication.

Authors' Contributions

Karthikeyan K: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Visualization, and Writing – Original Draft Preparation. Geetha B G: Methodology, Validation, Data Analysis, Visualization, and Writing – Review and Editing. Batumalay M: Supervision, Technical Guidance, Validation, Project Administration, and Writing – Review and Editing. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Acknowledgements

The authors have no specific acknowledgements to declare.

Code Availability

The source code developed for the proposed VIT-DRN-OCAFF framework is not currently publicly available. The implementation details required to reproduce the methodology are described in the manuscript. The corresponding author may provide additional implementation information upon reasonable request, subject to institutional and data-use restrictions.

Generative Artificial Intelligence and AI-Assisted Technologies

The authors declare that no generative artificial intelligence or AI-assisted technology was used to generate the scientific data, experimental results, statistical analyses, or scientific conclusions presented in this manuscript. The authors take full responsibility for the accuracy, originality, and integrity of the final manuscript.

Author Agreement

All authors have read and approved the final manuscript and agree with its submission to *Scientific Reports*. The authors confirm that the manuscript is original, has not been published previously, and is not currently under consideration for publication in another journal.

References

- [1] Mondal, A. K., Bhattacharjee, A., Singla, P., & Prathosh, A. P. (2021). xViTCOS: Explainable vision transformer-based COVID-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, 10, 1–10. <https://doi.org/10.1109/JTEHM.2021.3134096>
- [2] Shome, D., Kar, T., Mohanty, S. N., Tiwari, P., Muhammad, K., AlTameem, A., Zhang, Y., & Saudagar, A. K. J. (2021). COVID-transformer: Interpretable COVID-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21), 11086. <https://doi.org/10.3390/ijerph182111086>
- [3] Wu, Y., Qi, S., Sun, Y., Xia, S., Yao, Y., & Qian, W. (2021). A vision transformer for emphysema classification using CT images. *Physics in Medicine & Biology*, 66(24). <https://doi.org/10.1088/1361-6560/ac3dc8>
- [4] Mehboob, F., Rauf, A., Jiang, R., Saudagar, A. K. J., Malik, K. M., Khan, M. B., Hasnat, M. H. A., AlTameem, A., & AlKhathami, M. (2022). Towards robust diagnosis of COVID-19 using vision self-attention transformer. *Scientific Reports*, 12(1), 8922. <https://doi.org/10.1038/s41598-022-13039-x>
- [5] Jadhav, S., Deng, G., Zawin, M., & Kaufman, A. E. (2022). COVID-view: Diagnosis of COVID-19 using chest CT. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 227–237. <https://doi.org/10.1109/TVCG.2021.3114851>
- [6] Zheng, L., Fang, J., Tang, X., Li, H., Fan, J., Wang, T., Zhou, R., & Yan, Z. (2022). PVT-COV19D: Pyramid vision transformer for COVID-19 diagnosis. *arXiv Preprint arXiv:2206.15069*. <https://arxiv.org/abs/2206.15069>
- [7] Gao, X., Qian, Y., & Gao, A. (2021). COVID-ViT: Classification of COVID-19 from CT chest images based on vision transformer models. *arXiv Preprint arXiv:2107.01682*. <https://arxiv.org/abs/2107.01682>
- [8] Park, S., Kim, G., Oh, Y., Seo, J. B., Lee, S. M., Kim, J. H., Moon, S., Lim, J. K., & Ye, J. C. (2021). Vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. *arXiv Preprint arXiv:2104.07235*. <https://arxiv.org/abs/2104.07235>
- [9] Solomon, J. J., Heyman, B., Ko, J. P., Condos, R., & Lynch, D. A. (2021). CT of post-acute lung complications of COVID-19. *Radiology*, 301(2), E383–E395. <https://doi.org/10.1148/radiol.2021211396>
- [10] Hossain, R., Jeudy, J., & White, C. S. (2021). Chest radiographic and CT findings in patients hospitalized with breakthrough COVID-19. *Radiology: Cardiothoracic Imaging*, 3(6). <https://doi.org/10.1148/ryct.210248>
- [11] Goyal, S., & Singh, R. (2023). Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques. *Journal of Ambient Intelligence and Humanized Computing*, 14(4), 3239–3259. <https://doi.org/10.1007/s12652-021-03464-7>
- [12] Jasmine Pemeena Priyadarsini, M., Kotecha, K., Rajini, G. K., Hariharan, K., Utkarsh Raj, K., Bhargav Ram, K., ... & Pandya, S. (2023). Lung diseases detection using various deep learning algorithms. *Journal of healthcare engineering*, 2023(1), 3563696. <https://doi.org/10.1155/2023/3563696>

- [13] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., ... & Xu, B. (2021). A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *European radiology*, 31(8), 6096-6104. <https://doi.org/10.1007/s00330-021-07715-1>
- [14] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2021). Densely connected convolutional networks for medical image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7), 2261–2275. <https://doi.org/10.1109/TPAMI.2019.2918284>
- [15] Padmavathi, V., & Ganesan, K. (2025). LungNet-ViT: Efficient lung disease classification using a multistage vision transformer model from chest radiographs. *Journal of Vibration and Control*, 33(4). <https://doi.org/10.1177/08953996251320262>
- [16] Choi, H., Jin, S., & Han, K. (2025). ICEv2: Interpretability, comprehensiveness, and explainability in vision transformer. *International Journal of Computer Vision*, 133, 2487–2504. <https://doi.org/10.1007/s11263-025-02431-5>
- [17] Xiao, X., Kong, Y., Li, R., Wang, Z., & Lu, H. (2024). Transformer with convolution and graph-node co-embedding: An accurate and interpretable vision backbone for predicting gene expressions from local histopathological images. *Medical Image Analysis*, 91, 103040. <https://doi.org/10.1016/j.media.2023.103040>
- [18] Xu, H., & Wu, Y. (2024). G2ViT: Graph neural network-guided vision transformer enhanced network for retinal vessel and coronary angiograph segmentation. *Neural Networks*, 176, 106356. <https://doi.org/10.1016/j.neunet.2024.106356>
- [19] Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., & Wang, L. (2024). Vision transformers for image classification: A comparative survey. *Artificial Intelligence Review*, 57(2), 1–36. <https://doi.org/10.1007/s10462-023-10638-8>
- [20] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- [21] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [22] T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. <https://doi.org/10.1109/TIT.1967.1053964>
- [23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [24] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. <https://doi.org/10.1109/CVPR.2017.243>

NOMENCLAUTURE

CT	Computed Tomography
ViT	Vision Transformer
DRN	Dilated Residual Network
OCAFF	Optimized Cross-Attention Feature Fusion
ViT-DRN-OCAFF	Vision Transformer–Dilated Residual Network with Optimized Cross-Attention Feature Fusion
COVID-19	Coronavirus Disease 2019
CNN	Convolutional Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
RF	Random Forest
VGG16	Visual Geometry Group 16
ResNet50	Residual Network with 50 Layers
DenseNet121	Densely Connected Convolutional Network with 121 Layers
CLAHE	Contrast Limited Adaptive Histogram Equalization
HU	Hounsfield Unit
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
F1	F1-Score
NPV	Negative Predictive Value
MCC	Matthews Correlation Coefficient
GPU	Graphics Processing Unit
Adam	Adaptive Moment Estimation

Greek symbols

Q	Query Feature Matrix
K	Key Feature Matrix
V	Value Feature Matrix
d	Feature Dimensionality Scaling Factor
μ	Mean Pixel Intensity
σ	Standard Deviation of Pixel Values
I	Input / CT Image
α	Adaptive Attention Coefficient
β	Adaptive Attention Coefficient
HU	Hounsfield Unit
s	Spatial Scaling Factor
N	Number of Image Patches