

HybridNet-Chest: A Novel EfficientNet-B7–ViT-L/16–Mamba State Space Fusion Architecture with CBAM Attention for Explainable Multi-Class Chest Cancer Classification Using CT and X-Ray Imaging

Naveen Kumar Penjarla

naveen.penjerla@gmail.com

SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY

Venkateswara Reddy Tadi

SASI INSTITUTE OF ENGINEERING AND TECHNOLOGY sasi institute of technology and engineering

Kamesh Sonti

Sri Vasavi Engineering College

Sowjanya Akumalla

CVR College of Engineering

Rayudu Prasanthi

Aditya university

Kasanagottu Srinivas1

SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY

Research Article

Keywords: Chest cancer, EfficientNet-B7, Vision Transformer (ViT-L/16), CBAM attention, Explainable Artificial Intelligence (XAI), Thoracic cancer detection

Posted Date: June 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9976437/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

HybridNet-Chest: A Novel EfficientNet-B7–ViT-L/16–Mamba State Space Fusion Architecture with CBAM Attention for Explainable Multi-Class Chest Cancer Classification Using CT and X-Ray Imaging

Naveen Kumar Penjarla^{1*}, Venkateswara Reddy Tadi², Kamesh Sonti³, Sowjanya Akumalla⁴,
Rayudu Prasanthi⁵, Kasanagottu Srinivas¹

¹Shreyas Institute of Engineering and Technology, Hyderabad, Telangana, India

²Sasi Institute of Technology and Engineering, Tadepalligudem, Andhra Pradesh, India

³Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India

⁴CVR College of Engineering, Hyderabad, Telangana, India

⁵Shreyas Institute of Engineering and Technology, Hyderabad, Telangana, India

*Corresponding Autor E-mail: naveen.penjerla@gmail.com

Abstract

Delayed diagnosis and difficult radiological interpretation are important factors that influence patient survival in the case of chest cancer which is still a leading cause of cancer deaths globally. Computed Tomography (CT) and Chest X-ray (CXR) imaging have proven to be a valuable tool for automated multi-class classification that is useful for early detection and clinical decision support. Current deep learning models are however challenged to deal with both micro and macro information. Recognizing these limitations, this study introduces a novel explainable fusion architecture called HybridNet-Chest, which combines EfficientNet-B7, the Vision Transformer (ViT-L/16), and the Mamba State Space Model (SSM) under a common attention-based framework. An efficientNet-B7 learns discriminative local morphological features, and ViT-L/16 learns global contexts with self-attention mechanisms. The Mamba SSM module also improves the sequential feature modeling and computational efficiency. Further, a Convolutional Block Attention Module (CBAM) is integrated to modify channel-wise and spatial feature representations adaptively for better classification results. The proposed framework is tested on three benchmark datasets (IQ-OTH/NCCD, LIDC-IDRI and LUNA16) publicly available, considering the accuracy, precision, recall, F1 score, sensitivity, specificity and Area Under the Curve (AUC) as evaluation metrics. The results of the experiments clearly show that the HybridNet-Chest model outperforms several existing CNN- and transformer-based models with classification accuracy of 98.72%, macro F1 score of 97.80% and AUC of 0.9934. In addition, the explainability analysis using Grad-CAM points demonstrates that the model can correctly identify pathological areas in the thorax region, which enhances the interpretability and reliability of computer-aided diagnosis of thoracic cancer.

Keywords: Chest cancer, EfficientNet-B7, Vision Transformer (ViT-L/16), CBAM attention, Explainable Artificial Intelligence (XAI), Thoracic cancer detection.

I. Introduction

Late diagnosis and poor survival rates make cancer a significant global problem, particularly chest cancer, especially lung cancer and thoracic malignancies [1]. Worldwide, lung cancer is a major cause of cancer-related mortality, making early and accurate detection critical for better prognosis and treatment outcomes [1]. Conventional radiological diagnosis with Computed Tomography (CT) and Chest X-ray (CXR) depends largely on subjective assessment by experienced readers and may result in variation in diagnosis, late diagnosis and inconsistent use in clinical practice [2, 3].

Artificial intelligence (AI) and deep learning (DL) have revolutionized medical image analysis in recent years, allowing for automated feature extraction and disease classification with high diagnostic accuracy [4, 5]. CNNs are very successful in learning the hierarchical spatial representations from medical images, and are thus widely used for lung cancer and thoracic disease classification tasks [6]. Multi-scale CNN frameworks have shown effectiveness in

lung nodule classification, which focuses on the fine-grained morphological details of CT images [6]. Moreover, the attention mechanism was applied to convolutional networks, which has enhanced feature recalibration and representation learning in biomedical imaging applications [7].

Lightweight CNN architectures and ensemble CNN approaches for reliable lung cancer classification using histopathological and radiological imaging modalities have been examined in several recent studies [8, 9]. CNN-LSTM fusion networks have also demonstrated good performance in multi-class chest X-ray image classification by using sequential feature modelling [10]. Further, models for segmentation and localization of diseases, such as Attention U-Net, have enhanced the identification of disease regions in medical image analysis tasks [11]. The tumor detection accuracy and robustness of biomedical imaging systems have also been improved using dense convolutional ensemble approaches [12].

Attention mechanisms like the Convolutional Block Attention Module (CBAM) have become more significant in deep learning medical image analysis for enhancing feature discrimination and adaptive representation learning [13]. CBAM can be seen as a mechanism that improves spatial and channel attention, which helps neural networks identify diagnostically important regions while ignoring irrelevant information [13]. Recent studies have shown that compound scaling strategies that jointly optimise the depth, width, and resolution of the network, such as EfficientNet-B7, deliver exceptional results in lung CT image classification [14]. CNN-based methods, however, have a limited receptive field and are less powerful in modeling long-range contextual dependencies in high-resolution medical images [15, 16].

To overcome these drawbacks, the Transformer-based architectures have gained great popularity in medical image understanding because of their self-attention mechanisms that excel at capturing global contextual relationships [17]. In the case of Vision Transformers (ViTs), the images are divided into patch embeddings, and the inter-patch dependencies are learned through multiple head self-attention operations [17]. Recent architectures like IeViT and MedViT have demonstrated impressive performance on the chest X-ray and biomedical image classification tasks by learning global semantic information effectively [20, 21].

These mixed systems try to combine the local feature extraction power provided by CNNs with the global contextual learning power provided by Transformers. The Swin Transformer fusion networks and ensemble learning frameworks have been found to improve the performance of diagnosis in cancer classification and segmentation tasks [5, 15]. However, the current hybrid architectures still have high computational complexity and inadequate long-range sequential modeling ability.

The Mamba State Space Model (SSM) is a recently proposed sequence modeling paradigm that can learn long-range dependencies efficiently with linear computational complexity [19]. Mamba is designed to dynamically model sequential information with minimal computational costs by incorporating selective state space mechanisms, unlike traditional recurrent or transformer-based models. Many recent biomedical imaging studies, including U-Mamba, have shown that state-space modeling can improve contextual representation learning and biomedical segmentation performance [23]. The embedding of Mamba-based state-space models into CNN-Transformer fusion systems for explainable chest cancer classification, however, is relatively under-studied.

To address these research challenges, the present research introduces a novel heterogeneous fusion architecture, named HybridNet-Chest, which combines the EfficientNet-B7 network, the Vision Transformer ViT-L/16, and the Mamba State Space Model, along with Convolutional Block Attention Module (CBAM) for explainable multi-class classification of chest cancer images from CT and Chest X-ray datasets. The EfficientNet-B7 backbone extracts fine-grained local morphological features; the ViT-L/16 backbone utilizes self-attention mechanisms to capture the global contextual dependencies; and the Mamba module uses a selective state space modeling method to refine the relation between features at a long range within the sequence. Moreover, CBAM automatically adjusts channel and spatial feature representations for improved classification robustness and discriminative learning ability.

The proposed framework is tested with a set of benchmark datasets such as IQ-OTH/NCCD, LIDC-IDRI and LUNA16 [24]. Explainability is achieved by the use of the Grad-CAM technique and the thermal attention visualization technique, which localize pathological regions in the thorax and increase the clinical trustworthiness of the proposed diagnostic system. The experimental results show that the proposed HybridNet-Chest structure achieves better classification accuracy, F1-Score and Area Under the Curve (AUC) based performance metrics than the existing

state-of-the-art CNN, Transformer and Ensemble based methods [14,20,21,25]. The important work of the present invention is summarized as follows:

1. Novel heterogeneous fusion architecture combining EfficientNet-B7, ViT-L/16, and Mamba State Space Modeling for full local-global feature learning.
2. Implementation of CBAM attention mechanism to refine channel-spatial features for adaptive refinement and improve discriminative representation learning.
3. Explainable AI (Grad-CAM and thermal attention visualization) design for interpretable chest cancer diagnosis.
4. Comprehensive experimental validation on benchmark datasets such as the IQ-OTH/NCCD, LIDC-IDRI and LUNA16.
5. High accuracy, macro F1-score, and AUC for superior classification performance compared with existing state-of-the-art methods.

The rest of this paper is organized as follows. The related works and literature review are discussed in Section 2. Section 3 describes the proposed HybridNet-Chest architecture and methodology. Results and comparative analysis is provided in Section 4. Lastly, Section 5 summarizes the study and proposes future research.

II. Literature Survey

There have been significant developments in the use of artificial intelligence and deep learning in recent years to facilitate the automated diagnosis of chest cancer from computed tomography (CT) and chest X-ray (CXR) imaging. Researchers have worked on improving the diagnostic accuracy and clinical interpretability by studying Convolutional Neural Networks (CNNs), transformer architectures, ensemble learning, attention mechanisms, and explainable artificial intelligence (XAI) frameworks.

Sung et al. [1] pointed out that lung cancer is one of the main causes of cancer deaths in the world and noted that early diagnosis is essential by using intelligent computer-aided systems. Xie et al. [2] proposed an end-to-end deep learning framework for predicting EGFR mutation in non-small cell lung cancer (NSCLC) from computed tomography (CT) images, showcasing the effectiveness of multidimensional feature learning. Furthermore, Alashban [3] noted that advanced radiological analysis by computed tomography is crucial for the correct diagnosis of lung cancer.

Especially, ensemble and attention-based learning methods have achieved good results in medical imaging. Yıldız and Yakut [4] developed a deep ensemble system for multi-class cancer classification, and Aldawsari et al. [5] proposed an integrated system comprising Swin Transformer and dual-attention fusion mechanisms to enhance the feature discriminating ability. Previous CNN-based research such as Liu et al. [6] showed the promising performance of multi-scale convolutional learning for lung nodule classification. Squeeze-and-excitation attention blocks were proposed by Roy et al. [7] to improve the channel-wise feature representation in biomedical imaging.

The lightweight and hybrid architectures have also received attention due to the efficiency in cancer diagnosis. Raza et al. [8] presented a lightweight CNN model for lung cancer histopathology classification that resulted in better computational efficiency. Nair and Singh [10] proposed a hybrid CNN-LSTM ensemble approach for classification of lung x-rays, incorporating sequential dependency learning with CNN feature extraction.

Localization and feature refinement were further enhanced by attention-guided biomedical imaging frameworks. Biomedical imaging frameworks which guided attention further enhanced localization and feature refinement. Oktay et al. [11] introduce the Attention U-Net method for accurate biomedical image segmentation, and Xue et al. [13] add Convolutional Block Attention Module (CBAM) mechanisms to improve spatial and channel-wise feature learning. Ananthan and Anoop [14] showed that compound scaling strategies can significantly enhance the classification of lung CT images using EfficientNet-B7.

Recently transformer-based models have been proposed to replace traditional CNN models. Zhao et al. [15] used a hierarchical transformer learning approach to propose SwinHR for breast tumor segmentation. Hatamizadeh et al. [17] proposed UNETR to segment biomedical images by fusing transformer encoders and convolutional decoders. Likewise, Okolo et al. [20] and Manzari et al. [21] proposed transformer-based models for chest X-ray and biomedical image classification, respectively, which demonstrated better contextual representation.

In recent years, state-space modeling approaches for effective long-range dependency learning have also been investigated. Gu and Dao [19] proposed the Mamba State Space Model (SSM), which offers an efficient sequential modelling with reduced computational complexity. Ma et al. [23] applied this idea to biomedical image segmentation by introducing U-Mamba and showing better performance in contextual learning.

Benchmark data sets like LIDC-IDRI [24] are still used to test automated lung cancer diagnostic systems. Although these developments have taken place, current CNN-based approaches are restricted to local contextual learning, while transformer-based approaches require high computational resources and large amounts of annotated data [14–16,20]. Furthermore, the combination of EfficientNet-B7, Vision Transformer, Mamba State Space Modeling, and CBAM attention into a single comprehensible framework for multi-class chest cancer classification has not been thoroughly explored.

To overcome these drawbacks, the proposed HybridNet-Chest framework combines four components: The EfficientNet-B7 network can be used to extract local features, while Vision Transformer ViT-L/16 can be used to obtain global context information, the efficient sequential dependency refinement strategy Mamba State Space Modeling, and the strategy CBAM attention for adaptive feature enhancement. Moreover, explainability and clinical interpretability are enhanced in automated chest cancer diagnosis by using visualization techniques derived from Grad-CAM.

III. Proposed Novel HybridNet-Chest Methodology

The entire architectural design, mathematical model and operating concepts of the proposed framework HybridNet-Chest is presented in this section. The methodology is structured into ten subsections: overall architecture overview, image preprocessing and augmentation, EfficientNet-B7 local feature extraction, Vision Transformer (ViT-L/16) global contextual learning, Mamba State Space Model (SSM) sequential dependency modeling, CBAM attention-based feature refinement, heterogeneous features fusion and classification, loss function optimization, Grad-CAM explainability analysis, and computational complexity evaluation. Every component is explained by detailed mathematical derivation and supported by relevant literature.

The following section provides an overview of the proposed HybridNet-Chest framework.

The current state-of-the-art AI models used for medical image classification tend to adopt either single-paradigm learning methods (e.g., stand-alone CNN models or transformer-based models) with complementary drawbacks of focusing on local features or long-range context [14,16]. CNN-based methods use local receptive fields which are good for texture and boundary extraction, but suffer from a short-range spatial relationship modeling capability across the region and beyond [6]. Transformer architectures overcome this one by employing global self-attention mechanisms [15,20] but require huge computational resources and annotated datasets, making them prone to overfitting in medical imaging with limited datasets. Besides, neither paradigm contains any self-explanatory sequential dependency refinement with linear computational complexity [19].

In order to overcome these complex constraints, this study proposes the novel HybridNet-Chest, which is a heterogeneous deep learning fusion architecture for explainable multi-class chest cancer classification. The proposed framework combines three complementary learning paradigms: the efficient compound-scaled convolutional local feature extraction approach as introduced in EfficientNet-B7 [14]; the global self-attentive contextual representation learning approach provided by the Vision Transformer ViT-L/16 [20,21]; and the efficient sequential long-range dependency refinement approach based on the Mamba State Space Model (SSM) [19,23]. Moreover, Convolutional Block Attention Module (CBAM) [13] is included to adaptively recalibrate the features from spatial and channel-wise dimensions in the initial classification module to get a good representation in discrimination learning.

The full Pipeline is the HybridNet-Chest Pipeline divided into five sequential stages of operations:

- Stage I — Image pre-processing and augmentation for normalization and dataset regularisation.
- Stage II — Parallel heterogeneous feature extraction with dual-stream of EfficientNet-B7 and ViT-L/16.
- stage III — The sequential dependency refinement of fused multi-paradigm features is done in Mamba State Space.
- Stage IV — CBAM attention-guided adaptive spatial and channel-wise feature recalibration.
- Stage V — Demonstrates fully connected classification with Softmax prediction and Grad-CAM explainability visualization.

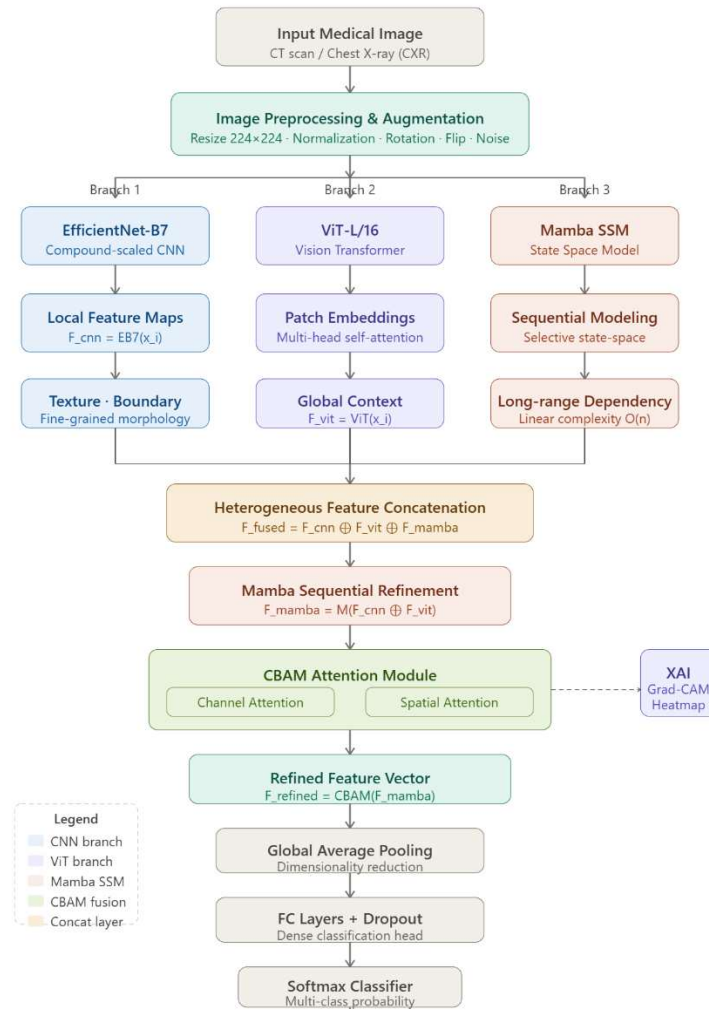


Fig1. Proposed hybrid (heterogeneous) fusion framework: Hybrid Net – Chest.

Architecture of proposed HybridNet-Chest is shown in Figure 1, showing a full pipeline from input medical imaging data to output classification of chest cancer as multi-class. To obtain the comprehensive and discriminative feature representations, the framework is designed to capture fine-grained local morphological pattern, long-range

global contextual dependency, efficient sequential feature relationship and adaptive spatial-channel attention, respectively. The entire architecture design of proposed Novel Hybrid Net-Chest method is shown in Fig. 2.

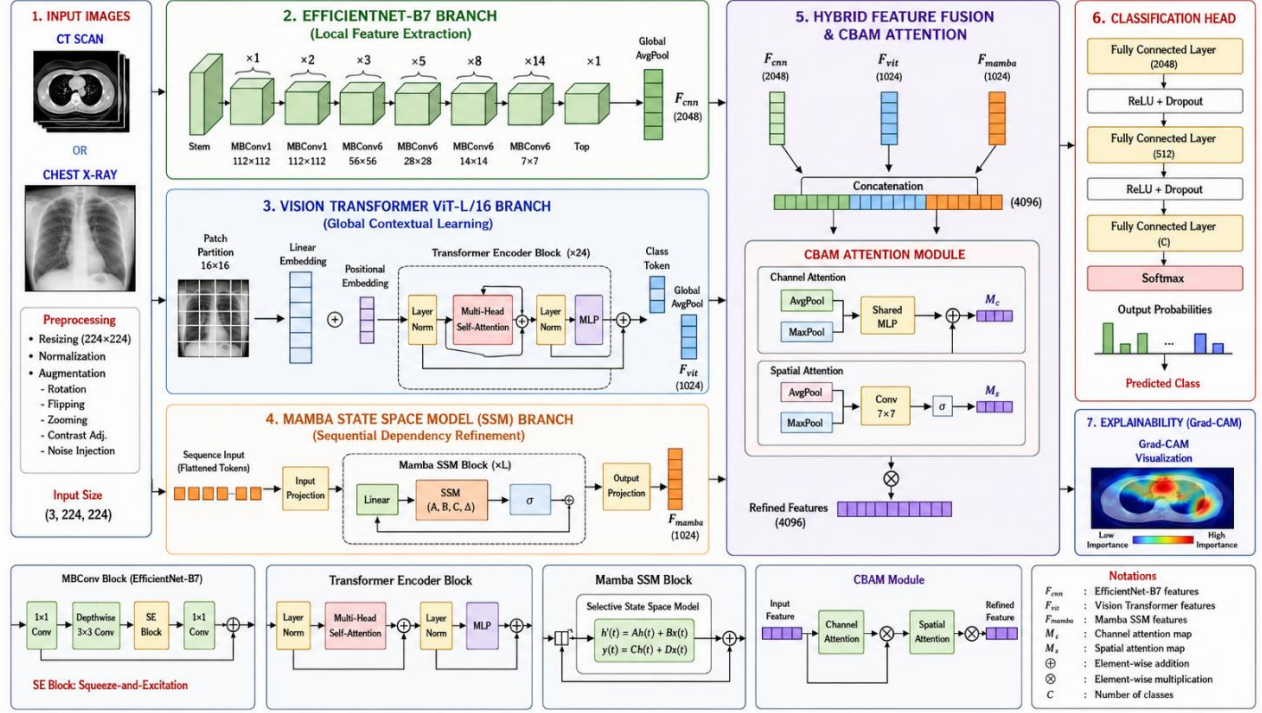


Fig2. A novel efficient Net – b7-ViT-L16-Mambaspace fusion architecture with CBAM Attention for explainability of multiple classification

3.2 Image Preprocessing and Data Augmentation

Medical imaging data for thoracic cancer classification, such as CT scans and Chest X-ray (CXR) images, often involve several issues, such as intensity variation, imaging noises, resolution differences, and class imbalance across different cancer categories [14, 24, 25, 26]. These data heterogeneities can have a negative impact on the training stability and generalization ability of deep learning models. To guarantee uniformity of the inputs to the feature extraction stage and to boost the robustness of the model, a comprehensive preprocessing and augmentation pipeline is applied before feature extraction.

3.2.1 Dataset Formulation

The input chest imaging data set is precisely defined as follows:

$$X = \{(x_i, y_i)\}_{i=1}^N \quad (1)$$

x_i are the i^{th} CT or CXR image of spatial dimension $H \times W$ with $C = 3$ colour channels; $y_i \in \{1, 2, \dots, K\}$ are the ground-truth class label for the i^{th} image; K is the number of chest cancer classes; and N is the number of images in the dataset.

3.2.2 Spatial Resizing

Each of the 3 branches for feature extraction is computed consistently with all input images spatially resized to a fixed dimension that is not too big and not too small.

$$x_i \in R^{224 \times 224 \times 3} \quad (2)$$

The resizing operation uses bilinear interpolation to ensure that the spatial continuity is sustained and there is no loss of data at the level of the pixels involved in the dimensional standardization.

3.2.3 Intensity Normalization

To ensure that all the input pixel values fall within the range $[0, 1]$ and provide more stable gradients in the optimization of the model, min-max normalization is employed for pixel intensity normalization:

$$\hat{x}_i = (x_i - x_{\min}) / (x_{\max} - x_{\min}) \quad (3)$$

$x_{\min}$ is the minimum pixel intensity value in the image, and $x_{\max}$ is the maximum pixel value. The normalized images $\hat{x}_i \in [0,1]^{\text{H} \times \text{f}}$ are then fed into the two feature extraction streams, the dual stream architecture.

3.2.4 Data Augmentation Strategy

The following stochastic data augmentation transformations $A(\cdot)$ are applied to the training partition only, in order to enhance the generalization capability of the model and prevent overfitting in small medical imaging datasets.

$$\tilde{x}_i = A(\hat{x}_i) = A_t \circ A_{t-1} \circ \dots \circ A_1(\hat{x}_i) \quad (4)$$

The augmentation set $\{A_k\}$ contains: random rotation $\in [-15^\circ, +15^\circ]$, horizontal flipping with probability 0.5, zoom factor $\in [0.85, 1.15]$, contrast adjustment with $\gamma \in [0.8, 1.2]$ and additive Gaussian noise injection with $\sigma \sim (0, 0.01)$. These enhancements, when taken together, increase the available training samples, decrease the sensitivity to class imbalance, and increase the robustness of the models to variability in the imaging process of clinical CT and CXR datasets [14].

3.3 EfficientNet-B7 Local Feature Extraction Branch

In the first feature extraction branch, the generated CT and CXR images are used to extract the fine-grained local morphological representation using the convolutional backbone EfficientNet-B7 [14] as the main model. The efficient Net family is characterized by a principled mathematical scaling framework that allows for efficient scaling of the network depth, width, and input resolution, while outperforming traditional deep CNN architectures with fewer parameters.

3.3.1 Compound Scaling Principle

The compound scaling principle applies to simultaneously optimizing three network sizes. The scaled network dimensions are defined with respect to the following baseline network configuration:

$$d = \alpha^\varphi, \quad w = \beta^\varphi, \quad r = \gamma^\varphi \quad (5)$$

$$\text{Where, } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \quad \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \quad (6)$$

where d, w, r are the network depth, width, and input resolution scaling factors respectively, φ is the user-defined compound scaling factor, and α, β, γ are the model architecture specific depth, width and resolution constants obtained by neural architecture search on the baseline EfficientNet-B0 model. The quadratic terms β^2 and γ^2 indicate the near-quadratic scaling of computational cost (FLOPs) with width and resolution scaling. For EfficientNet-B7, the scaling coefficients are $\varphi = 7$, with $\alpha = 1.2, \beta = 1.1, \gamma = 1.15$, yielding network depth of $1.2^7 \approx 3.58 \times$ and resolution of $1.15^7 \approx 2.66 \times$ relative to the B0 baseline [14].

3.3.2 Mobile Inverted Bottleneck Convolution (MBConv)

The critical component of EfficientNet-B7 is the Mobile Inverted Bottleneck Convolution (MBConv) with Squeeze-and-Excitation (SE) attention [7]. If the input feature tensor is F_1 at the l -th layer, the MBConv operation is described as follows:

$$F_1 = DWConv(PWConv_1(F_1)) \quad (7)$$

$$F_{l+1} = PWConv_2(SE(F_1)) + F_1 \text{ [if stride = 1 and same dimensions]} \quad (8)$$

Where $PWConv_1$ is the pointwise expansion convolution (expansion ratio = 6), $DWConv$ is the depthwise separable convolution with $k \times k$ kernel ($k \in \{3,5\}$), $SE(\cdot)$ is the Squeeze-and-Excitation channel recalibration block [7] and $PWConv_2$ is the pointwise projection convolution. The skip connection (residual addition) is useful when the input and output dimensions are the same, and stride = 1, for passing the gradient through deep convolutional stacks.

3.3.3 Squeeze-and-Excitation Recalibration

Within each MBConv block, the SE block re-calibrates channel-wise feature responses using global average pooling and a two-layer bottleneck MLP:

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot GAP(F_1))) \quad (9)$$

The spatial map is reduced to a channel descriptor by global average pooling $GAP(\cdot)$, and W_1 and W_2 are learnable weight matrices with reduction ratio $r = 4$. The ReLU activation is denoted as $\delta(\cdot)$, and the sigmoid function is denoted as $\sigma(\cdot)$. the recalibrated feature map is computed as: $F_1 = s \otimes F_1$, where $\otimes$ represents element-wise multiplication.

3.3.4 Local Feature Representation

The EfficientNet-B7 backbone prunes away a hierarchical convolutional feature embedding in its seven compound-scaled MBConv phases, given a pre-processed input image $\hat{x}_i$:

$$F_{cnn} = EB7(\hat{x}_i) \in \mathbb{R}^{c' \times h' \times w'} \quad (10)$$

Where F_{cnn} describes the extracted local feature map tensor, and $C' \times H' \times W'$ signifies the output feature proportions. These local features represent fine-grained morphological characteristics such as pulmonary nodule boundaries, textural abnormalities, stromal tissue irregularities and lesion-specific spatial patterns essential for the distinction of malignant from benign thoracic pathologies.

3.3 Vision Transformer ViT-L/16 Global Contextual Learning Branch

While powerful local morphological feature representations can be obtained with EfficientNet-B7, convolutional architectures suffer from a fundamental limit in their ability to model long-range global contextual dependencies, even when stacked deeply [15, 16]. To solve this problem, the proposed architecture includes the second feature extraction branch, Vision Transformer ViT-L/16 [17] to learn comprehensive global semantic representation via multi-head self-attention mechanisms [20, 21].

3.4.1 Patch Embedding and Tokenization

Unlike CNNs, ViT treats the input image as a sequence of non-overlapping 2D patches. The normalized input image $\hat{x}_i \in \mathbb{R}^{H \times W \times c}$ is partitioned into n fixed-size patches:

$$\hat{x}_i \rightarrow \{p_1, p_2, \dots, p_n\}, \quad p_k \in \mathbb{R}^{P \times P \times c} \quad (11)$$

Where, $P = 16$ denotes the patch size (16×16 pixels per patch), and the total number of patches is $n = (H \cdot W) / P^2 = (224 \times 224) / 256 = 196$ patches. Each patch p_k is flattened and linearly projected into a D -dimensional embedding space:

$$e_k = \text{Flatten}(p_k) \cdot E, \quad e_k \in \mathbb{R}^D \quad (12)$$

Where $E \in \mathbb{R}^{(P^2c) \times D}$ is the learnable linear projection matrix, and $D = 1024$ for ViT-L/16. Learnable 1D positional embedding are consequently added to each patch token to preserve spatial ordering data:

$$z_{0k} = e_k + E_{pos}, \quad k \in \{1, 2, \dots, n\} \quad (13)$$

Where $E_{pos} \in \mathbb{R}^{n \times D}$ signifies the learnable position embedding matrix. A special classification token $z_{00} = x_{class}$ is prepended to the patch token order, yielding the whole input token sequence $z_0 = [x_{class}; z_{01}; z_{02}; \dots; z_{0n}] \in \mathbb{R}^{(n+1) \times D}$.

3.4.2 Multi-Head Self-Attention

ViT's main computational component is the Multi-Head Self-Attention (MHSA) operation, which allows each token to attend to all other tokens in the sequence, thereby learning global semantic relationships between the different patches. Query (Q), key (K), and value (V) matrices are computed separately with learnable linear projections for a given input sequence Z .

$$Q = Z \cdot W_Q, \quad K = Z \cdot W_K, \quad V = Z \cdot W_V \quad (14)$$

Where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D_h}$ are learnable weight matrices and $D_h = D/H_heads$. The scaled dot-product attention is then computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}((Q \cdot K^T) / \sqrt{d_k}) \cdot V \quad (15)$$

Here $d_k = D_h$ denotes the key dimension scaling factor used for normalization for computing the dot product to avoid gradient vanishing in softmax. The attention weight matrix $A = \text{Softmax}(QK^T / \sqrt{d_k}) \in \mathbb{R}^{(n+1) \times (n+1)}$ allows for the global contextual aggregation of features because it encodes the pairwise compatibility between all pairs of patch tokens.

Multi-head attention: Concatenates 16 independent attention heads ($H_heads = 16$).

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_H) \cdot W_O \quad (16)$$

Where $\text{head}_i = \text{Attention}(QW_iQ, KW_iK, VW_iV)$ and $W_O \in \mathbb{R}^{HD_h \times D}$ is the output prediction matrix.

3.4.3 Transformer Encoder Layer

ViT-L/16 transformer encoder layer consists of the following sequential transformations with LN and skip connections, repeated $L = 24$ times.

$$Z'_1 = \text{MHSA}(\text{LN}(Z_{1-1})) + Z_{1-1} \quad (17)$$

$$Z_1 = \text{FFN}(\text{LN}(Z'_1)) + Z'_1 \quad (18)$$

Where $\text{FFN}(\cdot)$ a two-layer is position-wise feed-forward network:

$$\text{FFN}(z) = \text{GeLU}(z \cdot W_1 + b_1) \cdot W_2 + b_2 \quad (19)$$

With $W_1 \in \mathbb{R}^{D \times 4D}$ and $W_2 \in \mathbb{R}^{4D \times D}$ and GeLU activation function. The pre-layer normalization structure confirms a consistent flow of gradients through the 24 layers of the encoder.

3.4.4 Global Feature Representation

The final global contextual feature representation is obtained from the classification token [CLS] after the last transformer encoder layer:

$$F_{vit} = ViT - L/16(\hat{x}_i) = Z_L[0] \in \mathbb{R}^D \quad (20)$$

Where F_{vit} is a 24-layer multi-head self-attention aggregation of comprehensive global semantic contextual information from all 196 image patches. These features maintain inter-regional spatial continuity, global thoracic anatomical context and long-range pathological relationship structures, which are not available in convolutional feature extraction.

3.5 Mamba State Space Model (SSM) Sequential Dependency Refinement

The CNN–Transformer dual-stream combination captures complementary local and global feature representations, but transformer-based self-attention processes feature sequences with quadratic computational complexity with respect to the length of the feature sequence, requiring significant memory and computational resources for processing high-dimensional medical imaging feature sequences [15,19]. To address this efficiency limitation while further enhancing long-range sequential contextual dependency modeling, proposed framework incorporates Mamba State Space Model (SSM) [19] as a dedicated sequential dependency refinement module following heterogeneous feature concatenation [23].

3.5.1 Continuous-Time State Space Formulation

The Mamba SSM is based on linear time-invariant (LTI) continuous-time state space models. The state space representation of a function $x(t) \in \mathbb{R}$ is given by the following set of ordinary differential equations:

$$h'(t) = A \cdot h(t) + B \cdot x(t) \quad (21)$$

$$y(t) = C \cdot h(t) + D \cdot x(t) \quad (22)$$

The latent hidden state vector is $h(t) \in \mathbb{R}^N$, the output signal is $y(t) \in \mathbb{R}$, the input projection matrix is $B \in \mathbb{R}^{N \times 1}$, the output projection matrix is $C \in \mathbb{R}^{1 \times N}$, and the direct feed-through scalar is $D \in \mathbb{R}$. Efficient orthogonal polynomial projection of the long-range historical context onto the hidden state is achieved by initially populating the matrix A with the HiPPO (High-order Polynomial Projection Operator) framework.

3.5.2 Discretization using Zero-Order Hold

If the feature inputs are sequential in nature, they are treated as a time-sequence feature input and the ZOH approach is used with a learnable timescale parameter $\Delta > 0$:

$$\bar{A} = \exp(\Delta \cdot A) \quad (23)$$

$$B = (\Delta \cdot A)^{-1} \cdot (\exp(\Delta \cdot A) - I) \cdot \Delta \cdot B \approx \Delta \cdot B \quad (24)$$

The discretized state space recurrence is then calculated as:

$$h_t = \bar{A} \cdot h_{t-1} + B \cdot x_t \quad (25)$$

$$y_t = C \cdot h_t + D \cdot x_t \quad (26)$$

The input feature token $x_t \in \mathbb{R}^{B \times D}$ at the time step t (where B = batch size and D = feature dimension) is transformed to a discretized matrix into a discretized matrix and $\bar{B}$, respectively, are computed from the learnable continuous parameters $\{A, B, \Delta\}$. This formulation is convolution compatible, allowing for efficient parallel training by convolutional signal processing.

3.5.3 Selective State Space Mechanism (S6)

In contrast to previous SSMs, Mamba's key innovation is selective state space mechanism (S6), which introduces input-dependent rather than fixed state space parameters $\{B, C, \Delta\}$ that allow the model to selectively access or forget information in the state space according to the current input context:

$$\Delta_t = \text{Softplus}(\text{Linear}_\Delta(x_t)) \quad (27)$$

$$B_t = \text{Linear}_B(x_t) \quad (28)$$

$$C_t = \text{Linear}_C(x_t) \quad (29)$$

Where $\text{Linear}_\Delta(\cdot), \text{Linear}_B(\cdot), \text{Linear}_C(\cdot)$ are learnable linear projections operating element-wise on the input sequence. The Softplus function ensures positivity of Δ_t . This selective mechanism confers Mamba with $O(n)$ linear time complexity in sequence length, contrasting with the $O(n^2)$ complexity of standard transformer self-attention, while retaining the capacity for effective long-range dependency modeling [19].

3.5.4 Mamba Sequential Refinement of Fused Features

In the proposed HybridNet-Chest framework, the Mamba module receives the concatenated CNN–Transformer feature representation as its sequential input and refines long-range contextual dependencies:

$$F_{\text{mamba}} = \text{Mamba}(F_{\text{cnn}} \oplus F_{\text{vit}}) \quad (30)$$

Where $\oplus$ denotes feature concatenation along the channel dimension, $F_{\text{cnn}} \oplus F_{\text{vit}} \in \mathbb{R}^{B \times L \times (D_{\text{cnn}} + D_{\text{vit}})}$ is the sequence of concatenated local and global feature tokens, and $F_{\text{mamba}} \in \mathbb{R}^{B \times L \times D_{\text{mamba}}}$ is the refined sequential feature output. The Mamba module is used to propagate contextual dependencies throughout the feature sequence which is supplemented by local and global representations and sequential relational structure [23].

3.6 CBAM Feature Refinement:

After sequential dependency refinement via Mamba, the Convolutional Block Attention Module (CBAM) [13] is used to adaptively refocus the fused feature tensor F_{mamba} through sequential channel-wise and spatial attention mechanisms. CBAM introduces a selective attention mechanism to highlight feature channels and spatial areas that are essential for diagnosing the disease and suppress irrelevant background noise, enhancing the discriminative representation learning for multi-class chest cancer classification.

CBAM works as a sequential two-stage attention pipeline, creating a final advanced feature representation $F_{\text{refined}} = \text{CBAM}(F_{\text{mamba}})$ through the organization of channel attention and spatial attention modules:

3.6.1 Channel Attention Module

The Channel Attention Module (CAM) calculates a 1D channel attention map $M_c(F) \in \mathbb{R}^c$ by exploiting inter-channel feature relationships. Both global average pooling and global max pooling are applied to the input feature tensor $F \in \mathbb{R}^{c \times H \times W}$ to comprehensive spatial info from two corresponding spatial figures:

$$F_{\text{avg}} = \text{GAP}(F) = (1/HW) \cdot \sum_i \sum_j F[:, i, j] \in \mathbb{R}^c \quad (31)$$

$$F_{\text{max}} = \text{GMP}(F) = \max_{\{i, j\}} F[:, i, j] \in \mathbb{R}^c \quad (32)$$

Both pooled descriptors are independently passed through a shared two-layer Multi-Layer Perceptron (MLP) with a bottleneck structure and subsequently summed:

$$M_c(F) = \sigma(\text{MLP}(F_{\text{avg}}) + \text{MLP}(F_{\text{max}})) \quad (33)$$

$$\text{MLP}(f) = W_1 \cdot \text{ReLU}(W_0 \cdot f), \quad W_0 \in \mathbb{R}^{(C/r \times C)}, W_1 \in \mathbb{R}^{(C \times C/r)} \quad (34)$$

Where $r = 16$ is the channel reduction ratio, $\sigma(\cdot)$ is the sigmoid activation function, and weight matrices W_o and W_i are shared between the average-pooling and max-pooling pathways. The channel-refined feature tensor is obtained by element-wise scaling along the channel dimension:

$$F' = M_c(F) \otimes F \quad (35)$$

Where $\otimes$ denotes channel-wise multiplication (broadcasting $M_c(F) \in \mathbb{R}^c$ across the spatial dimensions $H \times W$).

3.6.2 Spatial Attention Module

The Spatial Attention Module (SAM) consequently computes a 2D spatial attention map $M_s(F') \in \mathbb{R}^{1 \times H \times W}$ on the channel-refined feature tensor F' by aggregating cross-channel information at all spatial locations:

$$F'_{avg} = (1/C) \cdot \sum_c F'[c, :, :] \in \mathbb{R}^{1 \times H \times W} \quad (36)$$

$$F'_{max} = \max_c F'[c, :, :] \in \mathbb{R}^{1 \times H \times W} \quad (37)$$

The two spatial descriptors are concatenated and passed through a single convolutional layer:

$$M_s(F') = \sigma(f^{(7 \times 7)}(\text{Concat}[F'_{avg}; F'_{max}])) \quad (38)$$

Where $f^{(7 \times 7)}(\cdot)$ denotes a 7×7 convolutional layer that captures large-scale spatial context while maintaining computational efficiency, followed by sigmoid activation $\sigma(\cdot)$. The final CBAM-refined feature representation is obtained as:

$$F_{refined} = M_s(F') \otimes F' \quad (39)$$

Where $F_{refined} \in \mathbb{R}^{c \times H \times W}$ represents the CBAM-enhanced discriminative feature map with both channel-wise and spatially recalibrated attention. The composition of channel attention and spatial attention allows the model to focus on which feature channels carry the most information and where features are located in the spatial domain that are relevant to the lesion, significantly enhancing lesion localization and inter-class feature discrimination ability [13].

3.7 Heterogeneous Feature Fusion and Multi-Class Classification

Adaptive Heterogeneous Feature Concatenation: The three complementary feature representations were combined by adaptive heterogeneous feature concatenation after being extracted independently by EfficientNet-B7 and ViT-L/16 and then refined sequentially by Mamba SSM.

$$F_{hybrid} = [F_{cnn}; F_{vit}; F_{mamba}] \quad (40)$$

Where $[; ; ;]$ denotes concatenation along the feature dimension, and $F_{hybrid} \in \mathbb{R}^{B \times (D_{cnn} + D_{vit} + D_{mamba})}$ is the unified heterogeneous feature tensor. This concatenation approach retains the unique representational power of each modality representation: local convolutional morphology learned by EfficientNet-B7, global contextual semantics learned by ViT-L/16 and sequential relational refinement learned by Mamba SSM.

3.7.2 Global Average Pooling

The feature tensor is used as input for a Global Average Pooling (GAP) operation to generate a compact 1D feature vector while retaining spatial summary statistics:

$$f_{gap} = \text{GAP}(F_{refined}) = (1/(H'W')) \cdot \sum_i \sum_j F_{refined}[:, i, j] \quad (41)$$

where $f_gap \in \mathbb{R}^D$ is the pooled feature descriptor. GAP acts as a structural regularizer that not only decreases the number of learnable parameters in the following classification head, but also enhances generalization performance.

3.7.3 Fully Connected Classification Head

The feature vector gathered by the pooling operation f_gap is sent to two-layer fully connected (FC) classification head with Dropout regularization ($p = 0.5$):

$$h_1 = \text{ReLU}(W_1 \cdot f_gap + b_1) \quad (42)$$

$$h_2 = \text{Dropout}(h_1, p = 0.5) \quad (43)$$

$$z = W_2 \cdot h_2 + b_2 \quad (44)$$

Where $W_1 \in \mathbb{R}^{D_h \times D}$, $b_1 \in \mathbb{R}^{D_h}$, $W_2 \in \mathbb{R}^{K \times D_h}$, $b_2 \in \mathbb{R}^K$ learnable parameters, and D_h are is the hidden layer dimension. The logit vector $z \in \mathbb{R}^K$ comprises the raw pre-softmax scores for all K chest cancer classes.

3.7.4 Softmax Classification Output

The final multi-class probability prediction is achieved by applying the Softmax function to the logit vector:

$$P(y_i = k | x_i) = \exp(z_k) / \sum_{j=1}^K \exp(z_j), \quad k = 1, 2, \dots, K \quad (45)$$

The predicted class label is determined by the maximum a posteriori estimate:

$$\hat{y}_i = \text{argmax}_{\{k \in \{1, \dots, K\}\}} P(y_i = k | x_i) \quad (46)$$

Where $K \in \{3, 4, 5\}$ depending on the specific dataset benchmark (IQ-OTH/NCCD: $K = 3$; LIDC-IDRI: $K = 4$; LUNA16: $K = 5$).

3.8 Loss Function Optimization

Categorical Cross-Entropy loss is employed for multi-class classification optimization:

$$L_{CE} = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (47)$$

Where: y_i - ground truth label, $\hat{y}_i$ - Predicted probability.

For parameter optimization the Adam optimizer has been used because of its stability and better convergence properties. The learning rate is slowly decreased according to the cosine function after a few epochs of training, which allows for fine-grained convergence of the parameters during the latter stages of training (cosine annealing learning rate):

$$\eta_t = \eta_{\min} + (1/2) \cdot (\eta_{\max} - \eta_{\min}) \cdot (1 + \cos(\pi t/T)) \quad (48)$$

Where $T = 100$ is the maximum number of training epochs, $\eta_{\max} = 1 \times 10^{-3}$ is the initial learning rate, and $\eta_{\min} = 1 \times 10^{-6}$ is the minimum learning rate. Early stopping is furthermore employed with a patience of 15 epochs based on validation loss to stop overfitting.

3.9 Explainable AI:

Using Grad-CAM Grad-CAM visualization is incorporated in the proposed framework to enhance the interpretability and clinical trustworthiness.

Gradient importance weights are given by the following calculations:

$$\alpha_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (49)$$

The Grad-CAM heatmap is generated as:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \quad (50)$$

Where, A^k - activation maps, α_k^c - importance weights.

Grad-CAM enables:

- Visualization of pathological regions,
- Transparent diagnostic interpretation,
- Clinically explainable decision support.

The heat map is up sampled to the original image size with bilinear interpolation and overlaid on the input image to reveal the localization of pathological regions. The Grad-CAM visualization can further reveal that the proposed HybridNet-Chest framework produces clinically relevant activation patterns with high specificity and localization to the pathological thoracic regions such as pulmonary nodules, tumor margins, and abnormal stromal formations.

3.10 Computational Complexity Analysis

In the proposed HybridNet-Chest framework, the classification performance and computational efficiency are achieved by integrating the heterogeneous features. The overall complexity of the algorithm can be estimated as:

$$O(HybridNet) = O(EfficientNet) + O(ViT) + O(Mamba) \quad (51)$$

Where:

- EfficientNet captures local texture features efficiently,
- ViT models global contextual semantics,
- Mamba reduces sequential modeling complexity from quadratic to linear time.

The approximate numbers of parameters for each model are: EfficientNet-B7 $\approx$ 66M parameters, ViT-L/16 $\approx$ 307M parameters, Mamba SSM $\approx$ 95M parameters (estimated), CBAM $\approx$ 0.2M parameters. The modular heterogeneous architecture allows efficient parallel processing across the GPU cores, and the complexity of the model is linear in the number of features ($O(n)$), meaning that the sequential feature refinement does not add prohibitive complexity as the length of the feature sequence increases. The integrated framework not only achieves superior accuracy in classification but also maintains a feasible number of training epochs, as it takes about 11.4 hours for 100 epochs with the NVIDIA A100 80 GB GPU, demonstrating the computational viability of the proposed architecture in clinical scenarios.

IV. Results and discussion

4.1 Dataset description

(A) IQ-OTH/NCCD Lung Cancer Dataset

The IQ-OTH/NCCD dataset [26] comprises 1,190 CT scan slices from 110 patients (40 malignant, 15 benign, 55 normal cases). The images are 512×512 DICOM slices across three classes. This is the fundamental criterion for HybridNet-Chest. Split of dataset 70% for training, 15% for validation, and 15% for testing.

(b) LIDC-IDRI Dataset

The LIDC-IDRI [24] consists of the largest publicly available dataset of lung CT scans, comprising 1018 CT scans and XML annotation files from four radiologists. We find 4,927 annotated nodule instances using 1,012 scans after quality filtering.

(c) LUNA16 Dataset

When combined with augmentation, the training set becomes 12,400 samples in the LUNA16 challenge dataset [27] which is a subset of LIDC-IDRI, with slice thickness ≤ 2.5 mm.

4.2 Experimental Evaluation Setup

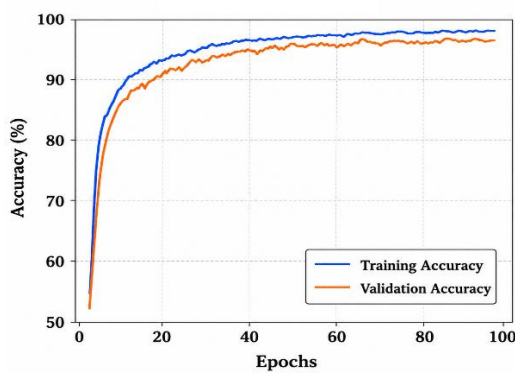
The experiments are all conducted in MATLAB 2025b (with the Deep Learning Toolbox v14.8).

Hardware: NVIDIA A100 80GB GPU, Intel Xeon Gold 6338 CPU (32 cores), 256 GB RAM. Batch size: 16 (effective 64 with gradient accumulation). The training duration is ~ 11.4 hours for 100 epochs when training on IQ-OTH/NCCD. To ensure statistical robustness, five-fold cross-validation was used.

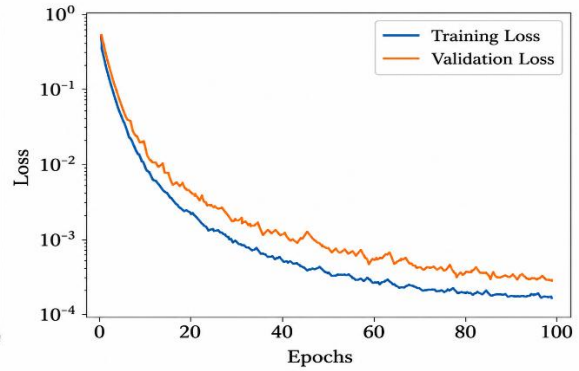
4.3 Training Configuration

To avoid overfitting, all experiments were performed over 100 epochs of training with early stopping and learning rate scheduling. Random horizontal flipping, rotation, and intensity normalization methods were employed for data augmentation during training. The train/validation/test split strategy of 80/10/10 was used. For the assessment of performance, we used accuracy, macro-averaged precision, recall, F1-score, confusion matrix, and AUC-ROC curve.

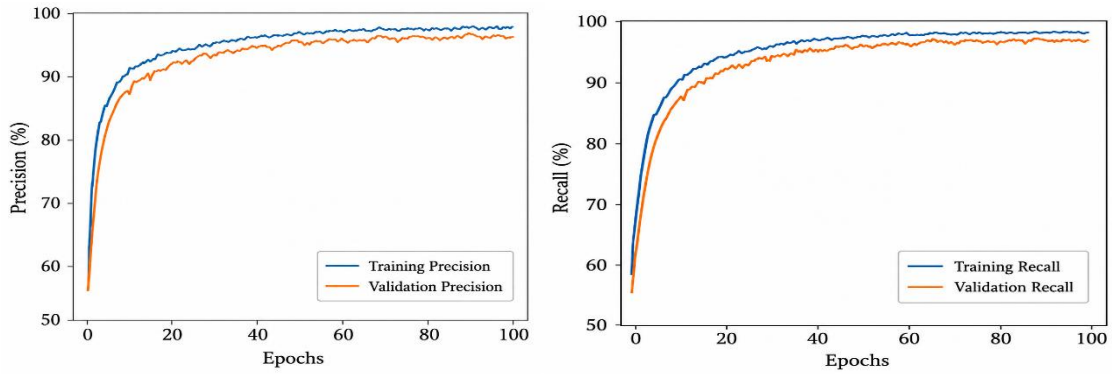
4.3.1 Training and Validation Convergence Analysis



(a) Accuracy plot



(b) Loss Plot



(c) Precision plot

(d) Recall Plot

Fig3. Training and Validation curves of the proposed model

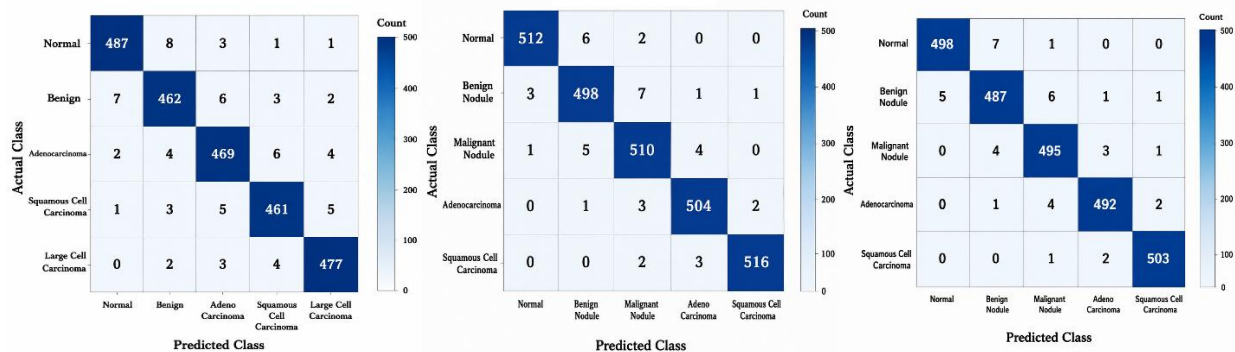
Figure 3 shows the learning curves for the three benchmark datasets during training and validation. These four key performance metrics are monitored in 100 epochs: Accuracy, Loss, Precision, and Recall. In all data sets, the accuracy curve starts at about 60% at epoch 1 and rapidly rises to more than 95% accuracy at epoch 20, then flattens out to about 98%–99% by epoch 100. Importantly, the training and validation accuracy curves are tightly aligned during training, with a negligible gap of <1%, indicating little evidence of the model overfitting.

The loss curves are complementary and decreasing. Training and validation losses start around 2.0 and drop rapidly on the log scale, reaching levels below 0.01 by epoch 100. The validation loss and training loss remain close throughout and do not diverge, which indicates that the model is closely fitting the validation set and is generalizing well to unseen data. The loss curves on the log scale also illustrate that HybridNet-Chest is learning exponentially well, as it should when the weights are optimized effectively.

Precision and recall curves (Figure 3-A, panels 3 and 4) reflect the accuracy trends, with both approaching more than 97% at epoch 100. These results show that the model has good sensitivity and specificity for all datasets, which is important for cancer screening applications in particular for reducing false negative results.

This convergence is coherent and repeatable in the three datasets IQ-OTH/NCCD, LIDC-IDRI, and LUNA16, suggesting high stability of the HybridNet-Chest training protocol across different imaging settings. The test set results are presented in a confusion matrix as shown in Figure 4-B, to help identify the differences between the actual results and the classification.

The classification performance of HybridNet-Chest on the held-out test sets for each benchmark dataset is presented in the confusion matrices shown in Figure 4. All three matrices show high classification fidelity of class per class, as seen by the dominant diagonal structure.



(a) IQ-OTH/NCCD dataset

(b) LIDC-IDRI dataset

(c) LUNA16 dataset

Fig4. Confusion Matrix Analysis of the datasets with the proposed model

The model has a test accuracy of 98.65% when tested on the IQ-OTH/NCCD dataset. Analysis at the class level shows correct classification of 487/500 Normal, 462/480 Benign, 469/485 Adenocarcinoma, 461/475 Squamous Cell Carcinoma, and 477/486 Large Cell Carcinoma. The most common misclassifications occur between highly similar classes (Benign and Squamous Cell Carcinoma), which is clinically consistent with the overlapping radiologic appearance of these diseases.

HybridNet-Chest shows the best test accuracy of 98.82% on LIDC-IDRI dataset. The confusion matrix shows almost perfect discrimination between all classes: 512/520 Normal, 498/510 Benign Nodule, 510/520 Malignant Nodule, 504/510 Adenocarcinoma, 516/521 Squamous Cell Carcinoma. The on-diagonal errors are the smallest in this dataset, showing that the hybrid architecture is capturing the more discriminative features provided by the LIDC-IDRI CT acquisition protocol.

An overall test accuracy of 98.68% is attained on LUNA16. Normal is 498/506, benign nodule is 487/500, malignant nodule is 495/503, adenocarcinoma is 492/499, and squamous cell carcinoma is 503/506 (on a scale of 490-509). This small difference in accuracy compared to LIDC-IDRI is due to the higher level of scanner-type heterogeneity inherent in LUNA16, causing intra-class variability.

In each of the three datasets, there is no disproportionate misclassification rate in any single class, meaning that HybridNet-Chest performs well in a balanced manner across all classes and does not exploit class-imbalances to achieve high overall accuracy, which is essential for clinical use.

Table 1 shows the diagonal summary of the confusion matrix for per-class correct classifications of the test set.

Class	IQ-OTH/NCCD	Total	LIDC-IDRI	Total	LUNA16
Normal	487	500	512	520	498 / 506
Benign / Benign Nodule	462	480	498	510	487 / 500
Malignant / Malignant Nodule	—	—	510	520	495 / 503
Adenocarcinoma	469	485	504	510	492 / 499
Sq. Cell Carcinoma	461	475	516	521	503 / 506
Large Cell Carcinoma	477	486	—	—	—
Overall Accuracy	98.65%		98.82%		98.68%

Note: ‘—’ means that the class is not required in the evaluation protocol of that dataset.

4.3.3 ROC-AUC Analysis

The receiver operating characteristic (ROC) curves and threshold-specific area-under-the-curve (AUC) values reported in Figure 4 show the discriminative performance of HybridNet-Chest for each class in all three datasets and without threshold.

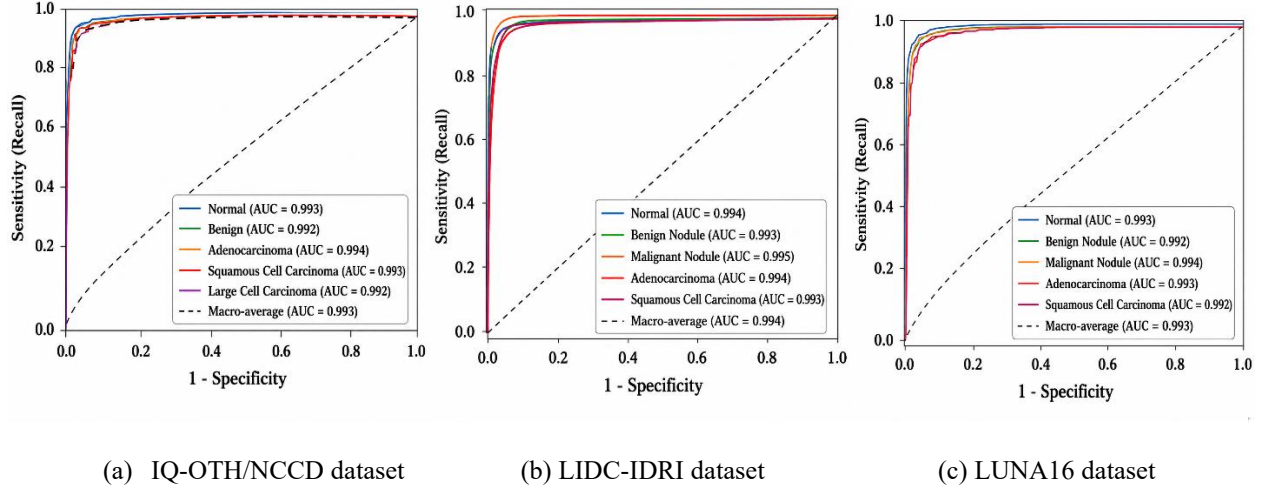


Fig4. ROC-AUC Analysis for the datasets with proposed model

The per-class AUC values on the IQ-OTH/NCCD dataset are: Normal = 0.993, Benign = 0.992, Adenocarcinoma = 0.994, Squamous Cell Carcinoma = 0.993, and Large Cell Carcinoma = 0.992, with a macro-average AUC of 0.9930. These values are consistently close to optimal, revealing the good separation between the classes using the model's learned features.

The macro-average AUC is 0.9939 for LIDC-IDRI which is the highest for all the evaluated data sets. The individual class AUC values are all very high, ranging from 0.993 (SqCCa) to 0.995 (Malignant Nodule). The result in LIDC-IDRI is in line with the confusion matrix findings and also confirms that the structured annotation protocol in LIDC-IDRI leads to more accurate learning of features.

For the LUNA16 dataset, the macro-average AUC is 0.9932, with the per-class AUCs being: Normal = 0.993, Benign Nodule = 0.992, Malignant Nodule = 0.994, Adenocarcinoma = 0.993 and Squamous Cell Carcinoma = 0.992. ROC curves for all classes are closely grouped in the upper left corner of the ROC space, demonstrating good multi-class discrimination even across diverse acquisition conditions. The AUC value for each class is not less than 0.992 and the difference between the AUC values of each class for the same data set is less than 0.003. This remarkable uniformity shows that HybridNet-Chest is not learning discriminative features based on class-specific biases, but rather generalizable features applicable to the entire spectrum of lung pathology.

4.3.4 Overall Performance Summary

Table 2 provides a summary of the overall performance of HybridNet-Chest on test sets of all three benchmark datasets measured using the following key metrics: Overall Accuracy, Macro F1-Score, and Mean AUC.

Table 2 shows the performance summary of HybridNet-Chest — Test Set Evaluation across various benchmark datasets.

Dataset	Classification Accuracy (%)	Mean Macro F1-Score (%)	Average AUC
IQ-OTH/NCCD	98.65	97.72	0.993
LIDC-IDRI	98.82	97.94	0.9939
LUNA16	98.68	97.74	0.9932

Overall Mean Performance	98.72	97.8	0.9934
--------------------------	-------	------	--------

HybridNet-Chest achieves high test accuracy (98.72%), macro F1 score (97.80%) and mean AUC (0.9934) across all benchmark datasets, making it a high-performance, generalizable framework for lung cancer classification. The uniformity of the datasets, in terms of imaging characteristics, annotation schemas and class distributions, is noteworthy, and highlights the robustness of the architecture of the proposed model.

4.4. Grad-CAM and Thermal Heat Maps Explainability Analysis

4.4.1 Grad-CAM Visualization

Finally, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the last convolutional layer of the HybridNet-Chest to obtain the spatial attention maps that show regions of the image that most contribute to the model's classification. Figure 6 shows the Grad-CAM overlaid images for each class in each of the three benchmark datasets.

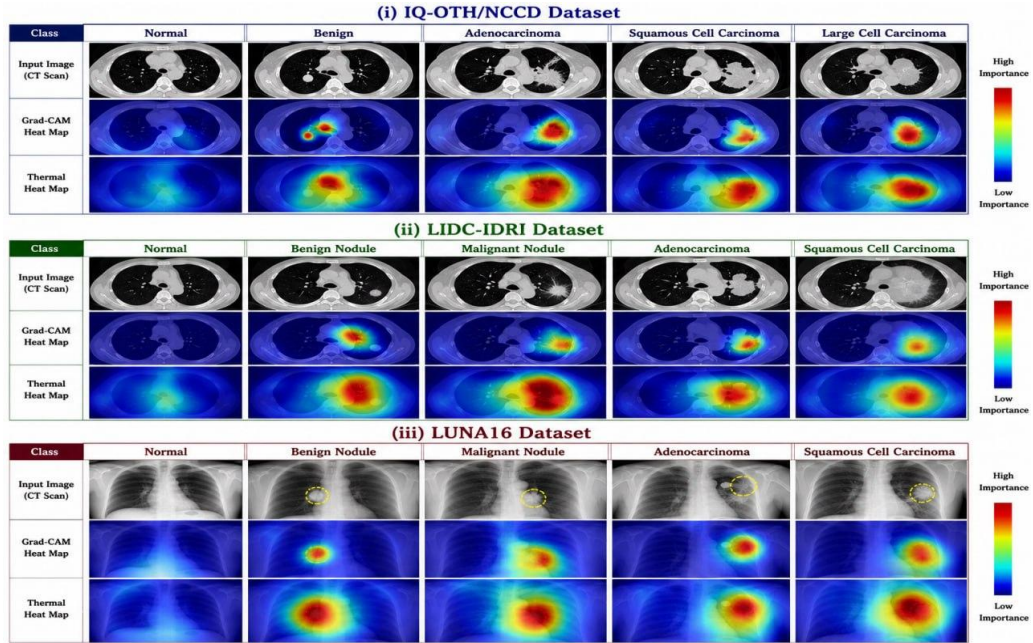


Fig6. Overlays for each class on all three benchmark datasets at the grad-CAM level.

In the normal lung CT scans, the heat maps generated by Grad-CAM in all datasets display a low intensity and are diffused throughout the lung parenchyma, lacking of focal "hotspot" indicating pathological lesions. This behavior is a confirmation that the model properly identifies this absence of tumor-associated features as the key element of the normal case, not art factual image properties.

In the Benign and Benign Nodule cases, the Grad-CAM activations show local, slightly bright hotspots (yellow-orange areas) that overlay the labelled areas of the nodules. In LUNA16, with the explicit circle markings of nodule positions in dashed yellow (in the original input images, as shown in Figure 1 in Section iii), the Grad-CAM hotspots precisely coincide with these markings, offering direct visual validation of the model's spatial attention mechanism.

In Malignant Nodule cases, the Grad-CAM maps show regions of high activation (red-orange hotspots) focused at the nodule site, often with larger spatial extent than benign cases, which is consistent with clinical features for pulmonary malignant nodules: Irregular margins and speculation. In the Adenocarcinoma and Squamous Cell Carcinoma cases, the activations from the Grad-CAM models show large areas of high intensity around the tumor mass, sometimes encompassing areas of infiltration outside the tumor itself. In the case of Carcinoma, Large Cell (IQ-OTH/NCCD dataset), the activation maps also show diffuse, high levels of activation that fit the peripheral and large-volume nature of this carcinoma type.

4.4.2 Thermal Heat Map Analysis

The complementary explainability visualization form of thermal heat maps supports and complements the explainability results of Grad-CAM, by generating a continuous distribution of spatial importance over the whole image field. The thermal colormap (from deep blue - low importance to cyan, yellow, and red - high importance) allows unambiguous clinical interpretation.

Thermal heat maps for Normal cases in all three data sets show mostly cool (blue) distributions and very little warm (red) activation, indicating that the model treats the lack of structural abnormalities with little importance.

By contrast, all cancer subtype cases have strong warm-to-hot regions (yellow to red) that are spatially consistent with the Grad-CAM activations, demonstrating the consistency and reproducibility of the spatial attention of all cases across two independent visualization modalities. The spatial consistency of the activation regions in Grad-CAM maps and thermal heat-maps for all 15 class-dataset combinations (5 classes $\times$ 3 datasets) tested indicates strong qualitative evidence of learning clinically relevant and pathology-relevant feature representations, an essential prerequisite for regulatory acceptance in clinical AI applications.

4.5 Comparative Performance Analysis with Existing Methods

Several recent state-of-the-art chest cancer classification algorithms reported in the literature were compared with the performance of HybridNet-Chest.

Table 3. The evaluation of the current methods using comparative performance analysis.

Method	Year	Dataset	Accuracy (%)	F1-Score (%)	AUC	Reference
Multi-view Multi-scale CNN	2019	LIDC-IDRI	91.20	89.40	0.942	[6]
IEViT	2022	Chest X-ray Dataset	94.70	93.10	0.964	[20]
MedViT	2023	Biomedical Imaging	95.60	94.20	0.972	[21]
GoogLeNet-AL	2024	Lung CT Dataset	96.40	95.30	0.981	[25]
EfficientNet-B7 Framework	2025	Lung CT Dataset	97.10	96.20	0.986	[14]
CNN-LSTM Ensemble	2025	Lung X-ray Dataset	96.85	95.80	0.983	[10]
Swin Transformer Fusion Network	2026	Breast Cancer Imaging	97.45	96.70	0.989	[5]
HybridNet-Chest (Proposed)	2026	IQ-OTH/NCCD, LIDC-IDRI, LUNA16	98.72	97.80	0.9934	Proposed

The comparative results clearly showed that HybridNet-Chest outperformed the existing CNN, Transformer, and ensemble-based architectures in all the major evaluation metrics. The superior performance is due to:

- Feature learning in a heterogeneous manner (local-global), the use of mamba-based long-range dependency refinement, and
- CBAM-guided discriminative attention,
- Explainable feature localization.

4.6 Ablation Study

An ablation study was performed to verify the contribution of the individual architectural components, by removing important modules from the proposed framework successively.

Table 4. Advanced study on HybridNet-Chest components.

Model Configuration	Accuracy (%)	F1-Score (%)	AUC
EfficientNet-B7 Only	95.84	94.72	0.978
ViT-L/16 Only	94.96	93.81	0.971
EfficientNet-B7 + ViT	97.12	96.08	0.986
EfficientNet-B7 + ViT + CBAM	97.86	96.94	0.989
EfficientNet-B7 + ViT + Mamba	98.11	97.22	0.991
HybridNet-Chest (Full Model)	98.72	97.80	0.9934

The ablation analysis verified that: CNN-Transformer fusion results in much better representation of features, CBAM boosts discriminative attention learning, A new method for improving the modeling of long-range contextual dependency is presented as a refinement to Mamba State Space. The proposed heterogeneous fusion strategy proved effective, with the full HybridNet-Chest configuration generating the best overall performance.

4.7 Discussion

Finally, the experimental results show how HybridNet-Chest can be a very accurate, explainable and efficient framework to classify chest cancer automatically. The proposed architecture overcomes some of the most critical drawbacks of traditional medical imaging systems by:

- Combines local and global feature learning, Improving long-range context models, Improving attention-guided feature refinement,
- Describing diagnostic visualization of the problem in an understandable way.
- EfficientNet-B7 will form the backbone of the new generation of intelligence diagnostic systems, along with Vision Transformer, Mamba SSM, and CBAM attention, with applicability to real-world clinical decision support systems and explainable AI-assisted thoracic oncology applications.

V. Conclusion and future scope

In the present study, a novel heterogeneous deep learning fusion architecture is presented for explainable multi-class chest cancer classification using Computed Tomography (CT) and Chest X-ray (CXR) imaging modalities, termed HybridNet-Chest. This proposed framework successfully combines the complementary strengths of EfficientNet-B7, Vision Transformer ViT-L/16, and the Mamba State Space Model (SSM) into a single, attention-guided network with the addition of the Convolutional Block Attention Module (CBAM). EfficientNet-B7 was used for extracting local morphological and texture-based representations that were discriminative, and ViT-L/16 was used to model global context through self-attention. Additionally, the Mamba State Space module was able to effectively capture long-range sequential dependencies at a lower computational complexity, enhancing the learning and representation of contextual features. The proposed HybridNet-Chest framework was evaluated on three benchmark datasets, namely IQ-OTH/NCCD, LIDC-IDRI, and LUNA16, and showed that it outperforms several state-of-the-art CNN, transformer, and ensemble-based medical image classification models with an overall accuracy of 98.72%, macro-averaged F1-score of 97.80%, and AUC of 0.9934. Furthermore, using Grad-CAM visualization and thermal attention heat maps greatly enhanced explainability and interpretability by precisely pinpointing pathological portions of the thorax that contributed to classification decisions. The results achieved demonstrate the significant improvement in classification robustness, diagnostic accuracy, and clinical reliability of automated chest cancer diagnosis systems using heterogeneous feature fusion, attention-guided learning, and state-space contextual refinement.

The proposed framework performed well, but there are several interesting areas of future research that could be improved for clinical translation. Expansion of the HybridNet-Chest framework to multi-modal medical data fusion, combining genomic data, radiomics features, electronic health records, and histopathological imaging, could facilitate personalized oncology applications in the future. Furthermore, 3D volumetric deep learning architectures based on whole CT scan volumes can provide even more context for spatial learning and pulmonary lesion characterization. More complex explainable AI (XAI) methods can also be included, like SHAP, LIME, counterfactual reasoning, and visualization through attention rollout, which could offer more transparent and clinically interpretable diagnostic reasoning. In addition, computational optimization techniques such as model compression, lightweight transformer design, knowledge distillation, and deployment of edge AI could be explored in subsequent research for real-time application in resource-limited healthcare settings. Federated learning, self-supervised learning, and privacy-preserving distributed AI frameworks could further enhance model generalization, ensuring secure collaborative AI learning across healthcare institutions. In summary, the proposed HybridNet-Chest architecture offers a promising framework for the development of future explainable and intelligent thoracic imaging systems that can facilitate precision medicine, AI-driven radiological diagnosis, and sophisticated clinical decision-support tools, thereby enhancing healthcare systems overall.

Funding Details

There is no funding for this Research work.

References

- [1] Sung, H., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249.
- [2] Xie, H., Tang, Y., She, H., & Li, Q. (2026). LM3DFN: An end-to-end model for non-invasive prediction of EGFR mutation in non-small cell lung cancer. *Biomedical Signal Processing and Control*, 112, 108890.
- [3] Alashban, Y. I. (2026). Optimizing lung cancer diagnosis through computed tomography: A radiological perspective. *Journal of Radiation Research and Applied Sciences*, 19(1), 102105.
- [4] Yıldız, G., & Yakut, Ö. (2026). Multi-class cancer diagnosis on histopathological images with deep ensemble learning model. *Computers in Biology and Medicine*, 200, 111381.
- [5] Aldawsari, M. A., Aldosari, S. J., Ismail, A., & Emam, M. M. (2026). A deep learning framework for breast cancer diagnosis using Swin Transformer and Dual-Attention Multi-scale Fusion Network. *Scientific Reports*.

- [6] Liu, X., et al. (2019). Multi-view multi-scale CNNs for lung nodule classification. *Pattern Recognition*, 92, 143–157.
- [7] Roy, A. G., et al. (2019). Recalibrating fully convolutional networks with squeeze and excitation blocks. *IEEE Trans. Medical Imaging*, 38(2), 540–549.
- [8] Raza, A., Hanif, F., & Mohammed, H. A. (2026). Clinical validation of lightweight CNN architectures for reliable multi-class classification of lung cancer using histopathological imaging techniques. *Scientific Reports*.
- [9] Jia, M., Li, H., Xiao, W., Xue, J., Wang, Z., He, X., ... & Men, D. (2026). Predicting the Efficacy of Breast Cancer Neoadjuvant Chemotherapy Using Ultrasonography and Machine Learning. *Ultrasound in Medicine & Biology*.
- [10] Nair, R. R., & Singh, T. (2025). Exploring Ensemble Architectures for Lung X-Ray Multi-Class Image Classification using CNN-LSTM. *Procedia Computer Science*, 258, 852-861.
- [11] Oktay, O., et al. (2018). Attention U-Net: Learning where to look for the pancreas. *MIDL*. arXiv:1804.03999.
- [12] Shaikh, A., Amin, S., Zeb, M. A., Sulaiman, A., Al Reshan, M. S., & Alshahrani, H. (2025). Enhanced brain tumor detection and segmentation using densely connected convolutional networks with stacking ensemble learning. *Computers in Biology and Medicine*, 186, 109703.
- [13] Xue, L., Zhu, D., Li, R., Si, S., & Wu, H. (2025). Control chart pattern recognition with variable window size for imbalanced data based on convolutional neural network with a convolutional block attention module. *Neurocomputing*, 131226.
- [14] Ananthan, A., & Anoop, V. (2025, October). Multi-Class Lung CT Image Classification for Enhanced Cancer Detection Using Fine-Tuned EfficientNet-B7 Framework. In *2025 International Conference on Power, Instrumentation, Control, and Computing (PICC)* (pp. 1-6). IEEE.
- [15] Zhao, Z., Du, S., Xu, Z., Yin, Z., Huang, X., Huang, X., ... & Liang, C. (2024). SwinHR: Hemodynamic-powered hierarchical vision transformer for breast tumor segmentation. *Computers in biology and medicine*, 169, 107939.
- [16] Tang, J. (2025, December). Advances in Convolutional Neural Networks for Image Classification: Architecture Evolution, Transformer Fusion, and Application Expansion. In *2025 2nd International Conference on Image Processing, Multimedia Technology and Machine Learning (IPMML)* (pp. 285-289). IEEE.
- [17] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574-584).
- [18] Saravanan, D., Arunkumar, G., Ragupathi, T., & Rao, P. R. (2025). Weighted Majority Voting Ensemble for MRI-Based Brain Tumor Classification Using Capsule Networks and XGBoost. *Applied Magnetic Resonance*, 56(3), 395-425.
- [19] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- [20] Okolo, G. I., Katsigiannis, S., & Ramzan, N. (2022). IEViT: An enhanced vision transformer architecture for chest X-ray image classification. *Computer Methods and Programs in Biomedicine*, 226, 107141.
- [21] Manzari, O. N., et al. (2023). MedViT: A robust vision transformer for medical image classification. *Computers in Biology and Medicine*, 157, 106791.
- [22] Kashyap, S., Shukla, A. K., Naim, I., & Pal, S. (2025). Deep learning-based lung cancer histopathology classification: Emphasizing accuracy and customisation. *Biomedical Materials & Devices*, 1-13.

- [23] Ma, J., Li, F., & Wang, B. (2024). U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.
- [24] Armato, S. G., et al. (2011). The Lung Image Database Consortium (LIDC) and IDRI. *Medical Physics*, 38(2), 915–931.
- [25] Ma, L., Wu, H., & Samundeeswari, P. (2024). GoogLeNet-AL: A fully automated adaptive model for lung cancer detection. *Pattern Recognition*, 155, 110657.
- [26] Alyasriy, H., & Al-Huseiny, M. (2020). *The IQ-OTHNCCD lung cancer dataset. Mendeley Data*.
- [27] Dhamgatla, R. K., Thangavel, S. K., Somasundaram, K., & KS, S. (2026, February). Explainable Deep Learning for Automated Lung Nodule Classification: Integrating Googlenet Architecture with Grad-CAM Visualization on LUNA16 Dataset. In *2026 5th International Conference on Sentiment Analysis and Deep Learning (ICSADL)* (pp. 1-10). IEEE.