Tsinghua Science and Technology
Doi: 10.26599/TST.2026.9010045

Research Article

MSCM-Net: Multi-scale CNN-Mamba Network for Pathological Complete Response Prediction of Lung Cancer

Jiancun Zhou, Hulin Kuang(✉), and Jianxin Wang

Abstract Accurate prediction of pathological complete response (pCR) is useful for clinical precision treatment of lung cancer. However, most existing pCR prediction methods are based on either convolutional neural networks (CNNs) or Transformers, which cannot effectively capture 3D global information or fuse features. Therefore, this study proposes a novel multi-scale CNN-Mamba network (MSCM-Net) to achieve accurate pCR prediction on CT scans of lung cancer patients. In each stage of the hybrid encoder, after channel splitting for parameter reduction, we design the CNN branch and Mamba branch to extract local and global features. Specifically, in the Mamba branch, we propose a novel intra-slice and inter-slice scanning mechanism to implement 8-way 3D scanning, thereby effectively capturing 3D global information. Furthermore, to better fuse CNN and Mamba features, we design a novel multi-scale aware feature fusion module with channel-level and multi-scale spatial level fusion. The proposed method is evaluated on a private dataset including 108 lung cancer patients who underwent neoadjuvant chemoimmunotherapy. Experimental results demonstrate that MSCM-Net achieves the best accuracy of 83.33% and an area under curve of 84.08%. Furthermore, results on the esophageal cancer and stroke prognosis prediction tasks validate the generalizability of MSCM-Net.

Keywords Pathological complete response prediction, Lung Cancer, Hybrid CNN and Mamba, Multi-scale Aware Feature Fusion

1 Introduction

In neoadjuvant chemoimmunotherapy, pathological complete response (pCR) is regarded as the gold standard for evaluating therapeutic efficacy and predicting long-term prognosis, which is defined as the absence of viable tumor cells in postoperative pathological specimens [1]. However, the assessment of pCR relies on postoperative histopathological examination, which is retrospective and invasive. Consequently, it cannot provide real-time guidance for clinical decision-making during treatment, thus hindering the dynamic optimization of personalized therapeutic strategies. Therefore, developing a noninvasive and accurate method to predict pCR after neoadjuvant chemoimmunotherapy for lung cancer based on pre-treatment CT imaging has become a pressing clinical need in the context of precision oncology.

Current studies on pCR prediction from CT imaging mainly fall into two categories. The first category is radiomics-based methods, which extract hand-crafted radiomic features to describe tumor phenotypes [2–6]. However, the effectiveness of these methods heavily depends on the quality and representativeness of the extracted features, which are often suboptimal. The second, more prevalent category comprises deep learning-based approaches, which automatically learn discriminative features in an end-to-end manner. Convolutional Neural Networks (CNNs), with their strong local feature extraction capability, excel at capturing textural and morphological details of tumors [7, 8]; Transformers leverage self-attention mechanisms to effectively model global dependencies [9–11]. Nevertheless, both architectures face inherent limitations: CNNs struggle to capture long-range contextual dependencies that are crucial for accurate pCR prediction, whereas Transformers suffer from quadratic computational and memory complexity when processing high-resolution 3D CT volumes.

Beyond pCR-specific studies, there are also some CNN and Transformer methods for medical image analysis that can be used as a reference [12–14],and some studies have designed hybrid networks of CNN and Transformer for medical image analysis [15–17]. Although such hybrids mitigate certain shortcomings, they still encounter challenges in achieving efficient global modeling of 3D volumetric data.

Recently, to maintain global modeling capacity while achieving linear computational complexity, State Space Models (SSMs) particularly the representative Mamba [18] have recently been introduced into medical image analysis. How-

• Jiancun Zhou is with College of Information and Electronic Engineering, Hunan City University, Yiyang, Hunan, China
• Jiancun Zhou, Hulin Kuang, and Jianxin Wang are with Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Hunan, China. E-mail: zhoujiancun@hncu.edu.cn, hulinkuang@csu.edu.cn, jxwang@mail.csu.edu.cn
(✉) To whom correspondence should be addressed.
Manuscript received: 2025-11-12; revised: 2026-03-13; accepted: 2026-04-30

ever, Mamba's unidirectional sequence-scanning mechanism is originally designed for its core application domain, natural language processing, and is inherently misaligned with the multi-dimensional spatial structure of medical images. Although existing studies attempt to adapt Mamba to two-dimensional images via block-wise or row–column scanning strategies [19, 20], such approaches tend to overlook critical inter-slice dependencies when extended to 3D CT data. Recent works [21–23] have begun to explore inter-slice scanning, yet their modeling of volumetric information remains incomplete.

Relying solely on either CNN or Mamba architectures cannot simultaneously capture local details and global dependencies, which may limit further improvement in pCR prediction performance. Therefore, constructing an effective hybrid CNN–Mamba architecture is highly desirable but remains challenging. The challenges primarily lie in: (1) within 3D lung CT scans, individual slices provide local cues such as tumor edge sharpness and internal heterogeneity, while inter-slice relationships reveal spatial distribution and volumetric changes; current Mamba scanning strategies fail to adequately model spatial dependencies across slices. Therefore, it is necessary to design a 3D Mamba scanning strategy that jointly models intra-slice details and inter-slice contextual information. (2) Simply concatenating or stacking CNN- and Mamba-derived features may introduce redundancy or interference, while existing feature fusion schemes often fail to effectively exploit their multi-scale characteristics, leading to suboptimal fused representations. Thus, a more refined feature fusion mechanism is required.

To address the aforementioned issues, this study proposes a MSCM-Net for predicting pCR after neoadjuvant chemoimmunotherapy in lung cancer. The core of this framework is a Mamba module specifically designed for 3D medical imaging, incorporating an Intra–Inter Slice Scanning (I^2S) mechanism. The proposed I^2S-Mamba consists of a token mixer and a channel mixer, where the token mixer performs eight-directional 3D scanning to enhance spatial information modeling. Furthermore, to achieve effective fusion between CNN and Mamba features, a Multi-scale Aware Feature Fusion (MAFF) module is introduced. The proposed approach is validated on a postoperative lung cancer pCR prediction dataset from Hunan Cancer Hospital and further evaluated on an esophageal cancer dataset undergoing the same neoadjuvant chemoimmunotherapy protocol. Experimental results demonstrate that our method achieves accurate pCR prediction with superior generalization capability, indicating its potential applicability to pCR prediction in other cancer types. The main contributions of this study are summarized as follows:

- We propose MSCM-Net for pCR prediction after neoadjuvant chemoimmunotherapy in lung cancer. The network integrates the local detail modeling of CNNs with the efficient global dependency modeling of Mamba, enhancing 3D CT feature representation.
- We design an innovative I^2S (Intra–Inter Slice) 3D Mamba scanning mechanism that combines intra-slice eight-directional scanning with inter-slice sequential modeling, thereby strengthening the model's ability to capture 3D spatial structures and tumor spatial distribution patterns.
- We introduce a MAFF module to enable effective multi-scale integration of CNN and Mamba features.
- Extensive experiments on real lung cancer datasets demonstrate that the proposed method achieves state-of-the-art performance in pCR prediction. Furthermore, cross-cancer validation on an esophageal cancer dataset confirms the generalizability and robustness of the proposed framework.

2 Methodology

2.1 Overview of the Proposed Method

The proposed Multi-scale hybrid CNN–Mamba Network (MSCM-Net) is illustrated in Fig. 1. The overall architecture consists of three major components: a convolutional stem, a CNN–Mamba hybrid encoder with four hybrid stages, and a final classification head.

The processing pipeline begins with the convolutional stem, which is composed of two $3 \times 3 \times 3$ convolutional layers designed to extract low-level features from the input CT volume. The core of the architecture lies in the hybrid encoder, where each stage integrates several key designs. First, a channel-splitting operation is applied to reduce computational complexity. The resulting feature maps are then processed by two parallel branches: a CNN branch and a Mamba branch. The CNN branch focuses on capturing local spatial details, while the Mamba branch models global contextual dependencies through a novel I^2S mechanism. The I^2S module, specifically designed for volumetric medical imaging, enables efficient and comprehensive modeling of 3D spatial information in CT scans. For clarity, the tensor dimensions at the key operations of the overall network are directly annotated in Fig. 1.

To integrate the complementary representations learned by the two branches, a MAFF module is introduced to effectively fuse multi-scale features from both CNN and Mamba pathways. Finally, the classification head takes the fused features

as input and generates the final pCR prediction.

2.2 CNN Branch

As illustrated in Fig. 1, at the beginning of each hybrid stage, a channel-splitting strategy equally divides the input feature maps along the channel dimension into two parts. Each part contains half of the original feature channels and is independently fed into the CNN and Mamba branches, respectively. This design not only effectively reduces the total number of model parameters but also ensures that the two branches focus on learning different types of representations, thereby minimizing potential feature interference.

Within the CNN branch, the basic processing unit consists of two stacked convolutional blocks. Each block follows the standard convolution normalization activation design pattern, comprising a 3D convolution layer, an instance normalization layer, and a LeakyReLU activation function in sequence. The first convolutional block performs an initial transformation to extract fundamental local spatial patterns, while the second block incorporates a residual connection to reuse features from the preceding block. This residual design alleviates the gradient vanishing problem in deep networks and enhances the robustness of feature representations. The process can be formulated as:

$$Y_1 = \text{LeakyReLU}(\text{InstanceNorm}(\text{Conv}_1(X))), \quad (1)$$

$$Y_2 = \text{LeakyReLU}(\text{InstanceNorm}(\text{Conv}_2(Y_1))) + Y_1, \quad (2)$$

where X denotes the input feature maps to the CNN branch, and Y_2 represents the final output. Here, LeakyReLU refers to the rectified linear unit with a negative slope, and InstanceNorm denotes the instance normalization operation. Through this concise yet effective design, the CNN branch is capable of capturing local morphological and textural features in CT images that are highly relevant to pCR prediction.

2.3 Mamba Branch Based on I^2S Mechanism

The overall structure of the Mamba branch mainly consists of two components: a token mixer and a channel mixer, as shown in Fig. 1. In the token mixer, the input features are first processed by layer normalization and a linear projection, followed by a depthwise separable convolution to capture local spatial patterns. A SiLU activation function is then applied to introduce nonlinearity. The processed features are subsequently fed into a SSM equipped with the proposed I^2S mechanism, and finally passed through another layer normalization and linear projection to effectively model 3D spatial information. The channel mixer, consistent with the design of Transformer architectures, is responsible for information interaction and integration along the channel dimension. The overall computation of the I^2S-based Mamba module can be formulated as:

$$Z_1 = \text{SiLU}(\text{DW}(\text{Linear}(\text{LayerNorm}(X)))), \quad (3)$$

$$Z_2 = I^2S(Z_1), \quad (4)$$

$$Z_3 = \text{Linear}(\text{LayerNorm}(Z_2)) + X, \quad (5)$$

$$Z_{\text{out}} = \text{MLP}(\text{LayerNorm}(Z_3)) + Z_2, \quad (6)$$

where X denotes the input to the Mamba branch and Z_{out} represents its final output. Here, I^2S refers to the intra–inter slice scanning operation, MLP denotes the multilayer perceptron, and DW indicates the depthwise separable convolution.

Since the original Mamba model was designed for one-dimensional sequences, its native unidirectional scanning mechanism struggles to effectively capture the rich 3D spatial structures present in CT volumes, leading to potential information loss. To address this limitation, we propose a SSM equipped with an I^2S mechanism, whose core structure is depicted in Fig. 2.

In this model, the input features are first processed by the I^2S mechanism, which performs eight directional 3D scans to comprehensively traverse the spatial domain. The outputs from all scanning directions are then aggregated via summation to form a unified representation encompassing complete spatial context. The implementation of a single 3D directional scan is illustrated in Fig. 3(a), which consists of two hierarchical levels: the intra-slice scan operates within each 2D slice following a top-left-to-bottom-right traversal order, while the inter-slice scan models dependencies across slices along the axial (Z) direction in a top-to-bottom order. A complete 3D scan thus integrates path information across three orthogonal planes: axial, sagittal, and coronal.

To achieve comprehensive spatial perception, an eight-directional scanning strategy is designed, as shown in Fig. 3(b). Specifically, four basic scanning patterns are first defined: two intra-slice traversal modes (row-first and column-first), each paired with two inter-slice traversal directions (through the sagittal or coronal plane). Then, by completely reversing the scan order of these four modes (from bottom-right to top-left intra-slice traversal and from bottom-to-top inter-slice traversal), four additional scanning paths are derived. Together, these eight scanning directions provide full spatial coverage, enabling the model to capture complex volumetric structures from multiple perspectives and thereby enhancing its ability to learn global 3D representations from CT images.

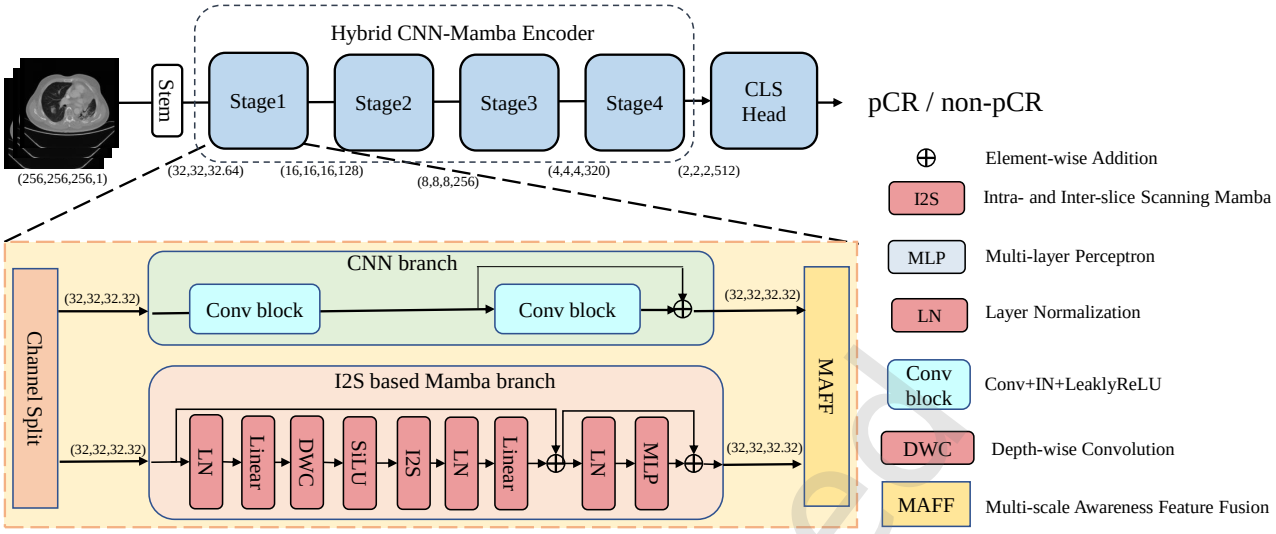

Fig. 1 Overview of the proposed MSCM-Net. The input and output feature sizes of Stem and each Stage are clarified.

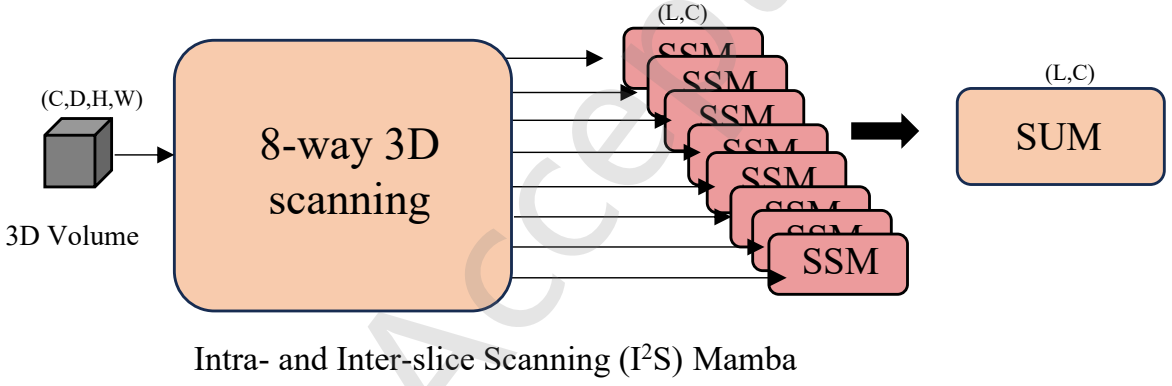

Fig. 2 Architecture of the SSM based on the proposed I^2S mechanism. The input 3D features are processed through eight directional 3D scans using the I^2S. Each directional scan is independently handled by a dedicated SSM, and the outputs of all eight paths are aggregated via summation to integrate spatial information from multiple perspectives.

2.4 Multi-scale Aware Feature Fusion Module

To effectively integrate the local details captured by the CNN branch with the global contextual representations modeled by the Mamba branch, a MAFF module is proposed, as illustrated in Fig. 4. During the fusion process, the features from the CNN and Mamba branches are first concatenated along the channel dimension. The concatenated features are then passed through a depthwise separable convolution followed by a pointwise convolution to achieve initial channel-level fusion.

However, relying solely on channel-level operations is insufficient to fully exploit the complementarity between the two types of features. To address this, a Multi-Head Mixed Convolution (MHMC) module is further designed to extract multi-scale spatial information. Specifically, the initially fused features are projected through a linear transformation and evenly divided into four subgroups along the channel dimension. Each subgroup is then processed in parallel by convolutional kernels of different sizes ($3\times3\times3$, $5\times5\times5$, $7\times7\times7$, and $9\times9\times9$), thereby capturing diverse information ranging from fine-grained local details to broader structural patterns. The parallel outputs are then concatenated and element-wise multiplied with another linearly transformed version of the initially fused features. This gating mechanism dynamically reweights and enhances discriminative feature responses. Finally, the features processed by the MHMC block are successively passed through batch normalization, pointwise convolution, and depthwise convolution layers to obtain more discriminative representations, followed by a residual connection and a final pointwise convolution to produce the fused output. For clarity, the tensor dimensions at the key operations of the MAFF module are directly annotated in Fig. 4. The overall fusion process of the MAFF module can

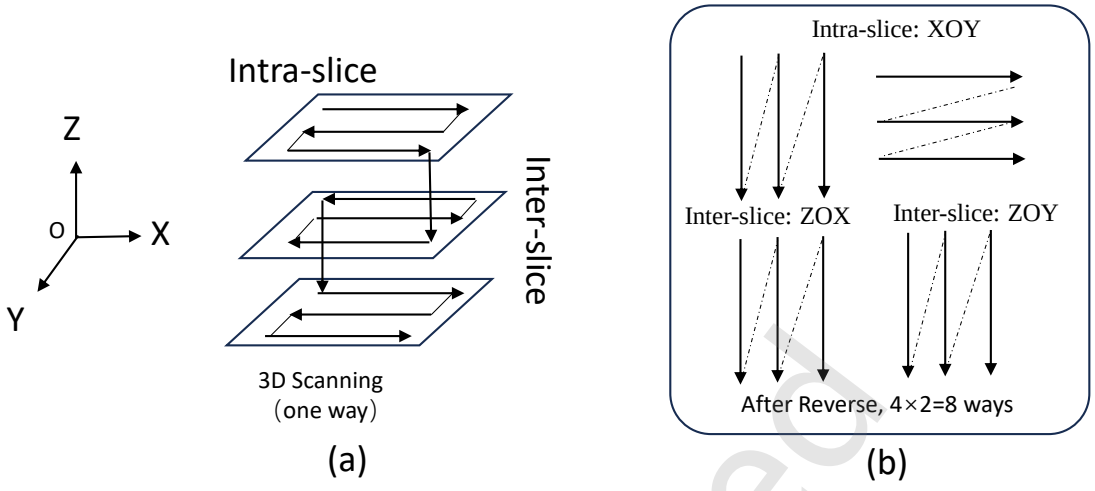

Fig. 3 Details of the eight-directional 3D scanning. (a) Illustration of a single directional 3D scan combining intra-slice scanning within each slice and inter-slice scanning across slices. (b) Generation of the eight scanning directions. By combining different scanning planes and reversing the scanning order, eight independent scanning paths are obtained, covering all spatial directions.

be expressed as:

$$F_{\text{hybrid}} = \text{Concat}(F_{\text{CNN}}, F_{\text{Mamba}}), \tag{7}$$

$$F_{\text{ch}} = \text{GELU}(\text{BN}(\text{PW}(\text{GELU}(\text{BN}(\text{DW}(F_{\text{hybrid}})))))), \tag{8}$$

$$F_{\text{sp}_k} = \text{Conv}_k(\text{ChannelSplit}(W_A \cdot F_{\text{ch}})), \tag{9}$$

$$F_{\text{MHMC}} = (W_B \cdot F_{\text{ch}}) \otimes \text{Concat}(F_{\text{sp}_k}), \tag{10}$$

$$F_{\text{fusion}} = \text{GELU}\Big(\text{BN}(\text{PW}(F_{\text{MHMC}})) + \text{BN}(\text{DW}(F_{\text{MHMC}})) + \text{BN}(F_{\text{MHMC}})\Big) \tag{11}$$

where F_{CNN} and F_{Mamba} denote the input features from the CNN and Mamba branches, respectively. Conv_k represents a convolution operation with a kernel size of $k \in \{3, 5, 7, 9\}$, used to extract spatial features at multiple scales. The symbol $\otimes$ denotes element-wise multiplication, which enables dynamic weighting and enhancement of key discriminative features. W_A and W_B correspond to the weight matrices of two independent linear projection layers. Note that for clarity, the normalization and activation layers embedded within the depthwise (DW) and pointwise (PW) convolution modules are omitted in the figure for simplicity.

2.5 Classification Head and Loss Function

The classification head designed in this study consists of a global average pooling (GAP) layer followed by a fully connected (linear) layer, which maps the encoded features to the final pCR classification prediction. Specifically, the 3D feature maps output by the encoder are first passed through the global average pooling layer to produce a 2D feature vector. This operation not only reduces the number of trainable parameters but also enhances the model's robustness to spatial variations in the input CT images. The resulting feature vector is then fed into a linear layer to generate the predicted probabilities of each sample belonging to the pCR-positive or pCR-negative class.

To effectively optimize model performance on the pCR prediction task, the cross-entropy loss function is employed as the training objective. This loss measures the divergence between the predicted probability distribution and the ground-truth label distribution. Its mathematical formulation is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}), \tag{12}$$

where N denotes the total number of samples in a batch, C is the number of classes, $y_{i,c}$ represents the ground-truth label, and $p_{i,c}$ denotes the predicted probability for class c of the i-th sample.

3 Experiments and Results

3.1 Datasets

As no publicly available dataset exists for predicting pCR after neoadjuvant chemoimmunotherapy in lung cancer, we collected a private dataset from Hunan Cancer Hospital for experimental validation. The dataset comprises 108 pathologically confirmed lung cancer patients. All patients underwent standard-dose non-contrast chest CT scans prior to the initiation of neoadjuvant chemoimmunotherapy, with scanning parameters standardized according to the hospital's imaging quality control protocols.

The determination of pCR labels was independently per-

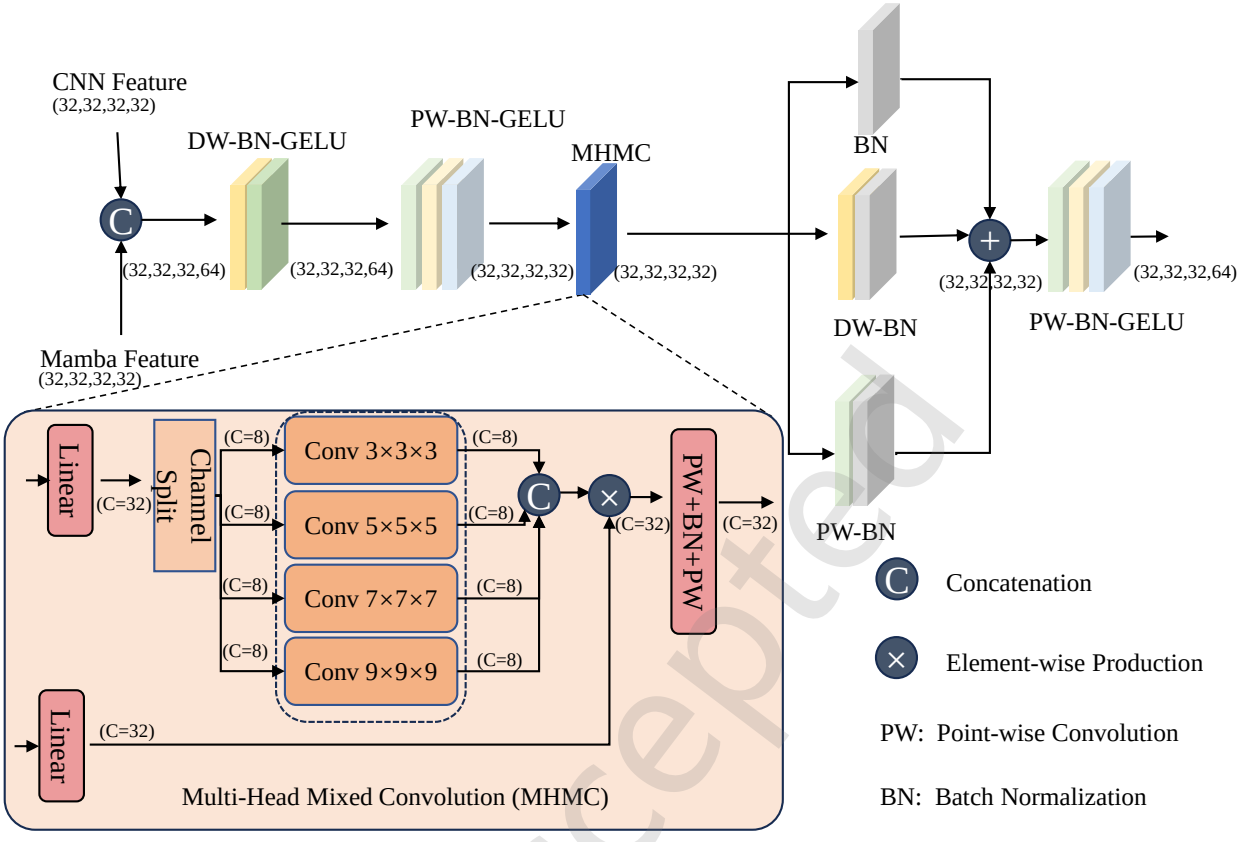

Fig. 4 Architecture of the proposed MAFF module. MAFF first concatenates the input features from the CNN and Mamba branches. The fused features are then processed by the core MHMC module, where the channel dimension is evenly split and fed into parallel convolutions with different kernel sizes. The annotated feature sizes are illustrated using Stage1 as an example.

formed by two experienced pathologists, who reached a consensus based on the postoperative pathological reports. Patients were labeled as pCR if no viable tumor cells were observed in the surgical specimens; otherwise, they were labeled as non-pCR.

As summarized in Table 1, the dataset consists of 108 patients, divided into the pCR group (46 cases, label = 1) and the non-pCR group (62 cases, label = 0). The two groups exhibited comparable demographic characteristics. In terms of age, the pCR group had an average age of 60.41 years, while the non-pCR group averaged 59.27 years. Regarding gender, both groups were predominantly male, accounting for approximately 60.87% (28/46) in the pCR group and 61.29% (38/62) in the non-pCR group. However, differences were observed in the pathological subtype distribution: the pCR group is dominated by squamous cell carcinoma (93.48%), whereas the non-pCR group showed greater heterogeneity, including both squamous cell carcinoma (66.13%) and adenocarcinoma (33.87%).

To further assess the generalization capability of the proposed model, an additional independent dataset comprising 243 esophageal cancer patients was collected from Sichuan Cancer Hospital. All patients received neoadjuvant chemoradiotherapy and underwent postoperative pathological evaluation. The pCR determination followed the same clinical criteria and expert consensus protocol, with pCR assigned as label = 1. Among these patients, 92 achieved pCR following neoadjuvant treatment, while 151 did not.

Table 1 Clinical characteristics of the lung cancer pCR prediction dataset.

	Label 0 (non-pCR)	Label 1 (pCR)
Number of samples	62	46
Mean age (years)	59.27 ± 7.09	60.41 ± 8.07
Gender (Female / Male)	24 / 38	18 / 28
Pathological subtype		
Squamous cell carcinoma	41 (66.13%)	43 (93.48%)
Adenocarcinoma	21 (33.87%)	3 (6.52%)

3.2 Experimental Settings

All experiments were conducted on an Ubuntu operating system using the PyTorch framework, and executed on an NVIDIA RTX 3090 GPU with 24 GB of memory. The model

was trained using the Adam optimizer with a batch size of 2 and 200 epochs. The initial learning rate was set to 1×10^{-4} and dynamically decayed using a polynomial learning rate scheduler with a decay factor of 0.9.

To alleviate overfitting, the following multiple strategies were jointly adopted. We applied weight decay in the optimizer and introduced stochastic depth to improve regularization and reduce parameter co-adaptation. An early stopping strategy based on the validation AUC was employed with a patience of 5 to prevent overfitting in the later stages of training. In addition, all experiments were conducted under a strict 5-fold cross-validation protocol to improve robustness and reduce potential bias caused by accidental data partitioning. Furthermore, several data augmentation techniques were employed during training, including random flipping, rotation, brightness adjustment, and Gaussian noise perturbation. Please refer to the code https://anonymous.4open.science/r/MSCM-Net-280E

3.3 Comparison Methods

To comprehensively evaluate the performance of the proposed method, MSCM-Net was quantitatively compared with state-of-the-art approaches on the collected lung cancer pCR prediction dataset. The results are summarized in Table 2. Due to the limited code availability of 3D medical image classification methods, all comparison models were faithfully reproduced following the technical details and publicly available implementations described in their original papers. The 3D version of DenseNet was implemented using the MONAI framework[①], while other 2D networks—including MedViT [24], RepViT [25], EViT [26], StarNet [27], UniRepLKNet [28], and FasterNet [29]—were adapted into their corresponding 3D versions for fair comparison. In addition, two deep-learning methods specifically designed for pathological pCR prediction were included for comparison, namely DELRN [30] and Voxel-Mamba [31]. DELRN is a multidimensional deep ensemble framework that integrates handcrafted radiomics features with deep features extracted from pretreatment CT images for pCR prediction, while Voxel-Mamba combines pretreatment CT images with voxel-level radiomics feature maps and performs classification using a 3D Vision-Mamba architecture.

Among the selected comparison methods, nnMamba [32] represents a hybrid CNN–Mamba architecture, while MedViT and EViT are Transformer-based frameworks. The remaining models are CNN-based architectures. For fairness, all methods were trained and evaluated using the same dataset splits and identical data augmentation strategies. Each baseline was reproduced using its official implementation, and all hyperparameters were re-tuned on the dataset used in this study to ensure optimal performance.

① https://monai.io/

3.4 Evaluation Metrics

To ensure fair comparison, all baseline methods were retrained and evaluated under the same experimental framework used in this study. We adopt five-fold cross-validation on the entire dataset. The evaluation metrics include accuracy (ACC), sensitivity (SEN), specificity (SPE), F1-score (F1), and the area under the ROC curve (AUC).

All reported performance values represent the mean and standard deviation calculated across the five validation folds.

3.5 Comparison of pCR Prediction Performance after Lung Cancer Treatment

Table 2 presents the comparison of the proposed method against eight state-of-the-art approaches on the lung cancer pCR prediction dataset after neoadjuvant chemoimmunotherapy. As shown, the proposed MSCM-Net achieved the best overall performance, with an ACC of 83.33%, SEN of 89.13%, F1 of 82.00%, and an AUC of 84.08%, outperforming all other competing methods across most metrics.

It is noteworthy that SEN is particularly critical for pCR prediction, as higher SEN helps minimize the risk of false negatives—ensuring that true pCR patients are not misclassified and thus avoiding the loss of potential opportunities for timely therapeutic adjustment.

Compared with nnMamba [32], which also adopts a CNN–Mamba hybrid architecture, MSCM-Net achieved a substantially higher AUC (84.08% vs. 74.33%), demonstrating the superiority of the proposed I^2S 3D scanning mechanism and MAFF module. Relative to Transformer-based models such as MedViT, the performance improvement of MSCM-Net can be attributed to the Mamba architecture's ability to efficiently model long-range dependencies in 3D volumes with linear computational complexity. In addition, MSCM-Net also outperformed pure CNN-based models such as StarNet, highlighting the importance of global contextual information in enhancing predictive accuracy for complex tasks like pCR prediction. Compared with DELRN [30], the superior performance of MSCM-Net suggests that the proposed framework can more effectively learn discriminative volumetric patterns in an end-to-end manner without relying on an explicit ensemble of handcrafted radiomics and deep features. Compared with Voxel-Mamba [31], which

Table 2 Comparison of pCR prediction performance on the lung cancer dataset. The best results are highlighted in bold. All values are reported as mean ± standard deviation.

Method	ACC(%)	SEN(%)	SPE(%)	F1(%)	AUC(%)	Parameters(M)	GPU memory(G)
nnMamba 2025 [32]	74.07 ± 2.68	76.09 ± 3.24	72.58 ± 3.63	71.43 ± 3.12	74.33 ± 2.87	13.62	13.78
MedViT 2023 [24]	78.70 ± 2.23	80.43 ± 3.15	77.42 ± 3.28	76.29 ± 2.51	78.93 ± 2.41	35.14	22.76
RepViT 2024 [25]	80.56 ± 1.94	78.26 ± 3.27	82.26 ± 3.05	77.42 ± 2.43	80.26 ± 2.17	**2.22**	4.90
EViT 2025 [26]	79.63 ± 2.14	67.39 ± 3.82	88.71 ± 2.36	73.81 ± 2.73	78.05 ± 2.52	26.16	3.64
StarNet 2024 [27]	80.56 ± 1.93	84.78 ± 2.94	77.42 ± 3.25	78.79 ± 2.35	81.10 ± 2.08	9.80	4.35
DenseNet 2017 [33]	70.37 ± 2.83	69.57 ± 3.91	70.97 ± 3.72	66.67 ± 3.34	70.27 ± 3.02	11.24	7.52
UniRepLKNet 2024 [28]	79.63 ± 2.16	80.43 ± 3.14	79.03 ± 3.18	77.08 ± 2.52	79.73 ± 2.43	43.46	10.33
FasterNet 2023 [29]	81.48 ± 1.82	69.57 ± 3.89	**90.32 ± 2.54**	76.19 ± 2.53	79.94 ± 2.31	18.02	**1.88**
DELRN 2025 [30]	80.56 ± 2.05	76.09 ± 3.12	83.87 ± 2.86	76.09 ± 2.64	81.12 ± 2.18	23.51	5.06
Voxel-Mamba 2025 [31]	81.48 ± 1.98	82.61 ± 2.88	80.65 ± 2.76	79.17 ± 2.24	82.50 ± 2.01	33.16	16.46
MSCM-Net (Proposed)	**83.33 ± 1.52**	**89.13 ± 2.03**	79.03 ± 2.14	**82.00 ± 1.83**	**84.08 ± 1.81**	16.54	7.48

already incorporates voxel-level radiomics priors and a modern Mamba-based backbone, MSCM-Net still achieves better results, indicating the advantage of the proposed hybrid CNN-Mamba encoder together with the MAFF-based multi-scale fusion strategy in capturing pCR-related information from pretreatment CT images.

3.6 Computational Complexity Analysis

We compare the number of trainable parameters and the peak GPU memory consumption of MSCM-Net and representative baselines under a unified evaluation setting in Table 2. MSCM-Net contains 16.54M trainable parameters and requires 7.48 GB peak GPU memory, indicating a moderate computational scale. Compared with a few generic baselines such as RepViT (2.22M, 4.90 G), MSCM-Net is moderately computational complex, but it remains obviously more efficient than most generic baselines such as MedViT (35.14M, 22.76 G) and also more economical than the pCR-specific baselines such as Voxel-Mamba (33.16M, 16.46 G). More importantly, MSCM-Net achieves the best overall pCR prediction performance in Table 2, with 83.33 ± 1.52% ACC and 84.08 ± 1.81% AUC. These results demonstrate that MSCM-Net achieves better trade-off between the computational complexity and predictive performance.

3.7 Ablation Studies

3.7.1 Effectiveness of the CNN Branch

To verify the contribution of the CNN branch, an ablation experiment was conducted by removing this component. As shown in Table 3, the overall ACC dropped by 9.26% after removing the CNN branch. The most notable decline occurred in SEN, which decreased by 19.56%, indicating a reduced ability to correctly identify pCR-positive patients. This result confirms that the local structural and textural features captured by the CNN branch—such as tumor boundaries and internal heterogeneity—provide valuable discriminative information for accurate pCR prediction.

3.7.2 Effectiveness of the I^2S-based Mamba Branch

To systematically assess the effectiveness of the Mamba branch with the proposed I^2S mechanism, three ablation experiments were conducted: (1) removing the Mamba branch entirely; (2) replacing the Mamba module with a Transformer block; and (3) substituting the proposed I^2S scanning mechanism with the original unidirectional Mamba scanning. The results (Table 3) show that removing the Mamba branch led to a 16.66% drop in ACC, highlighting its crucial role in modeling global features in CT volumes. Replacing Mamba with a Transformer resulted in a 4.63% ACC decrease, reflecting the higher efficiency and superior sequence-modeling capability of Mamba for long-range dependencies. Similarly, replacing the proposed I^2S with the original Mamba scanning mechanism also caused a 4.63% decline, validating the advantage of I^2S in effectively capturing 3D spatial context.

3.7.3 Effectiveness of the MAFF Module

To evaluate the contribution of the MAFF module, we replaced it with two common feature fusion strategies: direct concatenation and a cross-attention mechanism. As shown in Table 3, replacing MAFF with cross-attention reduced the ACC by 9.26%, while simple concatenation caused a more significant 12.96% decrease. These results demonstrate that the MAFF module, with its multi-scale perception and refined fusion design, more effectively integrates the CNN's local features and Mamba's global representations, leading to improved overall predictive performance.

3.7.4 Effectiveness of the Hybrid Architecture

To validate the necessity of the hybrid CNN–Mamba design, two additional experiments were conducted: (1) replacing both branches in the hybrid encoder with Mamba modules, and (2) replacing both branches with CNN modules. As presented in Table 3, using only Mamba in both branches led

Table 3 Ablation studies on the lung cancer pCR prediction dataset. A←B denotes replacing component A with component B. "CA" indicates cross-attention, and "Concat" denotes direct feature concatenation.

Methods	ACC(%)	SEN(%)	SPE(%)	F1(%)	AUC(%)
w/o CNN	74.07 ± 2.43	69.57 ± 3.52	77.42 ± 3.16	69.57 ± 3.24	73.49 ± 2.63
w/o Mamba	66.67 ± 2.91	78.26 ± 3.27	58.06 ± 3.85	71.11 ± 2.83	75.10 ± 2.54
Mamba←Transformer	78.70 ± 2.14	80.43 ± 2.94	77.42 ± 3.18	76.28 ± 2.45	78.92 ± 2.36
I^2S←Original Mamba Scanning	78.70 ± 2.15	88.24 ± 2.43	69.35 ± 3.47	78.50 ± 2.37	80.33 ± 2.24
MAFF←CA	74.07 ± 2.45	76.09 ± 3.28	72.58 ± 3.35	71.43 ± 2.86	74.33 ± 2.57
MAFF←Concat	70.37 ± 2.74	76.09 ± 3.25	66.13 ± 3.68	68.63 ± 3.07	71.11 ± 2.84
All branches use Mamba	71.30 ± 2.65	80.43 ± 3.08	64.52 ± 3.74	70.48 ± 2.94	72.48 ± 2.73
All branches use CNN	75.93 ± 2.37	88.24 ± 2.46	64.52 ± 3.75	76.36 ± 2.54	77.91 ± 2.43
MSCM-Net (Proposed)	**83.33 ± 1.52**	**89.13 ± 2.03**	**79.03 ± 2.54**	**82.00 ± 1.83**	**84.08 ± 1.81**

to a 12.03% drop in ACC, indicating that relying solely on global context is insufficient to capture the local imaging characteristics critical for pCR prediction. Conversely, using CNNs for both branches resulted in a 7.4% performance drop, revealing the limitations of local features when deprived of complementary global semantics. Together, these findings confirm that the proposed hybrid design effectively leverages the complementary strengths of CNN and Mamba, enabling simultaneous modeling of local details and global dependencies to enhance prediction performance.

4 Discussion

4.1 Interpretability Analysis

To investigate the decision-making rationale of the proposed model, feature activation visualization was conducted on axial CT slices from the test cohort. The visualizations were generated using Grad-CAM, and their purpose is to highlight the discriminative regions that contribute most to the model prediction. This approach visualizes the regions most strongly activated by the model, highlighting areas that contributed most to its final prediction. The resulting activation maps were superimposed on the original CT images in the form of heatmaps.

Fig. 5 presents four representative cases, including both relatively less satisfactory and comparatively better localization examples. As illustrated in Fig. 5, the activation regions with the highest values predominantly concentrate within or around the solid components of the tumor and its adjacent structures—particularly along the tumor core and boundary areas. These regions are consistent with previously reported imaging biomarkers that correlate with therapeutic response, including both intratumoral and peritumoral features [34]. Overall, these visualizations suggest that the proposed model tends to attend to lesion-related and clinically relevant regions.

4.2 Subgroup Analysis

To further evaluate the model's predictive performance across different patient populations, subgroup analyses were conducted on the lung cancer pCR prediction dataset according to three clinical variables: gender, pathological subtype, and age. Following the same evaluation procedure as the overall experiment, five-fold cross-validation was employed, and subgroup-specific metrics were computed for each fold. The final results were reported as the mean and standard deviation across all folds.

4.2.1 Gender-based Subgroup Analysis

As shown in Table 4, the female subgroup slightly outperformed the male subgroup across all metrics. Specifically, females achieved an ACC of 83.50%, SEN of 89.50%, SPE of 79.50%, F1 of 82.50%, and AUC of 84.50%, compared with 83.22%, 88.89%, 78.73%, 81.68%, and 83.81% for males, respectively. This marginally superior performance among female patients aligns with findings from previous studies suggesting gender-related differences in lung cancer treatment response [10].

Table 4 Comparison of model performance by gender (mean ± standard deviation). The better-performing subgroup is highlighted in bold.

Metric	Female (n=42)	Male (n=66)
ACC (%)	**83.50 ± 2.0**	83.22 ± 1.8
SEN (%)	**89.50 ± 2.5**	88.89 ± 2.2
SPE (%)	**79.50 ± 3.0**	78.73 ± 2.6
F1 (%)	**82.50 ± 2.2**	81.68 ± 2.0
AUC (%)	**84.50 ± 2.2**	83.81 ± 2.0

4.2.2 Pathological Subtype-based Subgroup Analysis

As presented in Table 5, the model achieved slightly higher performance on adenocarcinoma cases compared to squamous cell carcinoma. The ACC, SEN, SPE, F1, and AUC were 83.80%, 89.60%, 79.50%, 82.47%, and 84.55% for

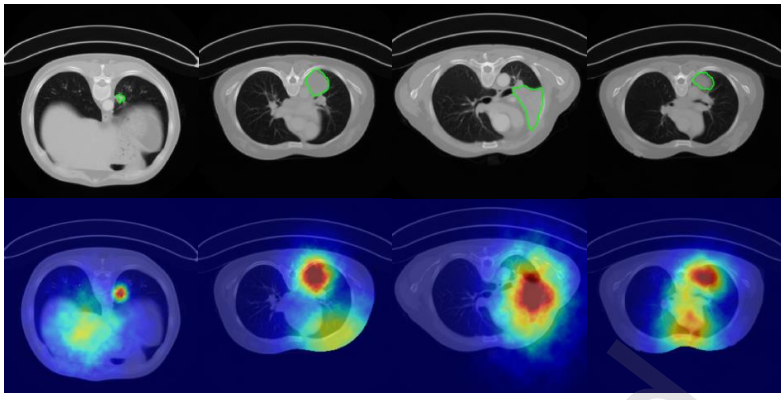

Fig. 5 Interpretability analysis of the proposed method. The first row shows the manually delineated ROI regions. The second row presents the Grad-CAM visualization results.

adenocarcinoma, versus 83.20%, 89.00%, 78.90%, 81.87%, and 83.95% for squamous cell carcinoma, respectively. These results suggest that the model demonstrated slightly better discriminative ability and positive detection performance in adenocarcinoma patients. The simultaneous improvement in F1 and AUC further indicates more stable classification behavior, which is consistent with prior CT-based response prediction research [10].

Table 5 Comparison of model performance by pathological subtype (mean ± standard deviation). The better-performing subgroup is highlighted in bold.

Metric	Squamous (n=84)	Adenocarcinoma (n=24)
ACC (%)	83.20 ± 1.8	**83.80 ± 3.0**
SEN (%)	89.00 ± 2.2	**89.60 ± 3.5**
SPE (%)	78.90 ± 2.6	**79.50 ± 3.2**
F1 (%)	81.87 ± 2.0	**82.47 ± 3.0**
AUC (%)	83.95 ± 2.0	**84.55 ± 3.0**

4.2.3 Age-based Subgroup Analysis

The age-based subgroup results are summarized in Table 6. The elderly group (>60 years) outperformed the younger group (⩽60 years) across all metrics, achieving an ACC of 83.50% vs. 83.16%, SEN of 89.50% vs. 88.76%, SPE of 79.50% vs. 78.56%, F1 of 82.50% vs. 81.50%, and AUC of 84.50% vs. 83.66%. These findings suggest that the model exhibited slightly better predictive stability in older patients, which aligns with previous clinical observations on age-related variability in treatment response [35].

4.3 Generalization to pCR Prediction after Neoadjuvant Chemoimmunotherapy in Esophageal Cancer

To validate the generalization capability of MSCM-Net on other cancer types for pCR prediction following neoadjuvant chemoimmunotherapy, an additional dataset was collected

Table 6 Comparison of model performance by age group (mean ± standard deviation). The better-performing subgroup is highlighted in bold.

Metric	>60 years (n=54)	⩽60 years (n=54)
ACC (%)	**83.50 ± 1.8**	83.16 ± 1.9
SEN (%)	**89.50 ± 2.0**	88.76 ± 2.1
SPE (%)	**79.50 ± 2.6**	78.56 ± 2.8
F1 (%)	**82.50 ± 2.1**	81.50 ± 2.3
AUC (%)	**84.50 ± 2.1**	83.66 ± 2.2

from Sichuan Cancer Hospital. This dataset comprises 243 patients diagnosed with locally advanced esophageal cancer, including both CT imaging and clinical information. For all patients, the pCR status was determined based on postoperative pathological examination, serving as the gold standard.

As in the lung cancer experiments, we adopt five-fold cross validation and report mean ± standard deviation. The esophageal cancer dataset was chosen for generalization testing due to the lack of publicly available CT-based lung cancer pCR datasets after neoadjuvant chemoimmunotherapy and the extended acquisition cycle of our private lung cancer dataset.

As summarized in Table 7, the proposed MSCM-Net achieved the best overall performance on the esophageal cancer pCR prediction task, with an ACC of 77.24%, SEN of 65.67%, SPE of 79.90%, F1 of 70.45%, and an AUC of 79.39%. These results demonstrate that MSCM-Net exhibits strong adaptability and robust generalization across different tumor types.

This generalization ability may be related to two aspects of MSCM-Net. On the one hand, tumor regions in 3D CT volumes usually exhibit a certain degree of inter-slice continuity, and the proposed I^2S mechanism explicitly models such spatial continuity and long-range dependencies among slices,

Table 7 Comparison of pCR prediction performance after neoadjuvant chemoimmunotherapy in esophageal cancer. The best results are highlighted in bold.

Methods	ACC(%)	SEN(%)	SPE(%)	F1(%)	AUC(%)
nnMamba 2025 [32]	72.96 ± 2.58	61.63 ± 3.63	73.42 ± 3.18	66.24 ± 2.97	71.07 ± 2.74
StarNet 2024 [27]	75.50 ± 2.25	63.81 ± 3.64	75.89 ± 2.81	68.97 ± 2.56	78.47 ± 2.09
EViT 2025 [26]	74.47 ± 2.29	62.60 ± 3.77	74.02 ± 3.12	67.01 ± 2.75	78.53 ± 2.31
UniRepLKNet 2024 [28]	74.99 ± 2.34	58.09 ± 3.72	76.54 ± 3.24	64.88 ± 2.89	72.77 ± 2.55
FasterNet 2023 [29]	76.74 ± 2.05	61.78 ± 3.69	78.83 ± 3.05	67.12 ± 2.46	77.99 ± 2.34
MSCM-Net (Proposed)	**77.24 ± 1.78**	**65.67 ± 3.28**	**79.90 ± 2.81**	**70.45 ± 2.31**	**79.39 ± 1.87**

Table 8 Comparison with SOTA methods for the AIS mRS prediction task.

Methods	ACC(%)	SEN(%)	AUC(%)
nnMamba 2025 [32]	74.62 ± 2.43	61.38 ± 3.72	73.95 ± 2.65
StarNet 2024 [27]	78.78 ± 2.14	63.08 ± 3.58	81.67 ± 2.18
EViT 2025 [26]	75.64 ± 2.37	61.96 ± 3.69	81.34 ± 2.23
UniRepLKNet 2024 [28]	76.98 ± 2.26	57.98 ± 3.85	75.01 ± 2.61
FasterNet 2023 [29]	78.66 ± 2.16	61.07 ± 3.73	80.24 ± 2.28
MSCM-Net (Proposed)	**82.12 ± 1.87**	**65.21 ± 3.42**	**83.19 ± 1.94**

which is helpful for extracting more globally informative representations from volumetric tumor regions. On the other hand, the MAFF module effectively integrates the local lesion details captured by the CNN branch with the global contextual information modeled by the Mamba branch, which may facilitate the learning of treatment-response-related tumor and peritumoral features. These characteristics may together contribute to the favorable generalization performance of MSCM-Net across different cancer types.

4.4 Generalization on the Stroke Prognosis Prediction task

To evaluate the generalization capability of MSCM-Net across different prognosis prediction tasks, we further collected a dataset comprising 295 patients diagnosed with acute ischemic stroke (AIS) and undergone mechanical thrombectomy from Xiangya Hospital, Central South University. The dataset comprises both imaging data and clinical information. The modified Rankin Scale (mRS) scores after 90 days for all patients were independently reviewed and confirmed by experienced radiologists and neurologists to ensure label reliability.

In this experiment, several representative state-of-the-art (SOTA) methods were compared with MSCM-Net for the AIS mRS prediction task to assess the model's cross-disease and cross-task applicability. As shown in Table 8, the proposed MSCM-Net achieved the highest performance, with an ACC of 82.12%, SEN of 65.21%, and an AUC of 83.19%. Its overall performance consistently outperformed all compared SOTA methods.

These results not only validate the effectiveness of MSCM-Net in stroke prognosis prediction but also demonstrate that the proposed hybrid CNN–Mamba architecture and the multi-scale fusion mechanism possess strong generalization capability. This indicates that MSCM-Net has promising potential for broader applications in post-treatment outcome prediction tasks across different diseases.

4.5 Effect of Kernel Sizes in the Multi-Head Mixed Convolution Module

To investigate the influence of kernel sizes in the MHMC module and to determine the most suitable multi-scale receptive field configuration for lung cancer pCR prediction, four kernel size combinations were evaluated: {3, 5}, {3, 5, 7}, {3, 5, 7, 9}, and {3, 5, 7, 9, 11}.

As shown in Table 9, the results demonstrate a clear "rise-then-drop" performance trend. When expanding the kernel combination from {3, 5} to {3, 5, 7} and further to {3, 5, 7, 9}, the model's AUC improved progressively from 77.49% to 80.50% and finally to 84.08%. This consistent increase indicates that incorporating a wider range of kernel scales enables the model to capture richer contextual information from lung CT images, facilitating multi-level spatial understanding of tumor structures. This result may be related to the complementary roles of different kernel scales. Specifically, the smaller kernels (3 and 5) mainly focus on local and fine-grained spatial patterns, which are helpful for capturing detailed lesion-related information, whereas the larger kernels (7 and 9) provide a broader receptive field that aggregates wider contextual information and long-range spatial dependencies. The {3, 5, 7, 9} configuration offers complementary modeling of local details and broader context, resulting in the better performance. However, when the 11×11×11 kernel was added, the AUC dropped to 80.01%, suggesting that an excessively large receptive field may introduce background noise or redundant contextual information unrelated to the tumor region, thereby distracting the model's focus from key discriminative features.

Table 9 Effect of different multi-scale kernel size configurations in the Multi-Head Mixed Convolution module.

Kernel Sizes	ACC(%)	SEN(%)	SPE(%)	F1(%)	AUC(%)
{3, 5}	79.63 ± 2.14	63.04 ± 3.82	**91.94 ± 2.36**	72.50 ± 2.73	77.49 ± 2.52
{3, 5, 7}	81.48 ± 1.82	73.91 ± 3.42	87.10 ± 2.68	77.27 ± 2.43	80.50 ± 2.18
{3, 5, 7, 9} (Proposed)	**83.33 ± 1.52**	**89.13 ± 2.03**	79.03 ± 2.54	**82.00 ± 1.83**	**84.08 ± 1.81**
{3, 5, 7, 9, 11}	79.63 ± 2.16	82.61 ± 2.76	77.42 ± 3.18	77.55 ± 2.45	80.01 ± 2.37

Table 10 Effect of different scanning strategies in the Mamba branch.

Scanning Strategies	ACC(%)	SEN(%)	SPE(%)	F1(%)	AUC(%)
Without flipped reverse	77.78 ± 2.14	76.09 ± 3.28	79.03 ± 3.18	74.47 ± 2.83	77.56 ± 2.52
Without inter-slice	76.85 ± 2.23	86.96 ± 2.68	69.35 ± 3.47	76.19 ± 2.53	78.16 ± 2.43
I^2S (Proposed)	**83.33 ± 1.52**	**89.13 ± 2.03**	**79.03 ± 2.54**	**82.00 ± 1.83**	**84.08 ± 1.81**

In summary, these results confirm that the {3, 5, 7, 9} kernel combination provides the optimal trade-off between multi-scale representation and noise suppression. It effectively captures hierarchical contextual cues while avoiding performance degradation caused by overly large receptive fields, making it the most suitable configuration for pCR prediction.

4.6 Effect of Scanning Strategies in the Mamba Branch

In the ablation studies, the proposed I^2S scanning mechanism demonstrated superior performance compared to the original Mamba scanning strategy. This improvement primarily stems from the comprehensive eight-directional scanning design, which enables more effective modeling of complex spatial structures in 3D medical images. To further examine the specific effects of different scanning strategies, two simplified variants were evaluated: (1) removing the flipped reverse scanning, and (2) removing the inter-slice scanning.

As shown in Table 10, when the flipped reverse scanning was removed, the model's AUC dropped to 77.56%. This finding indicates that multi-directional scanning serves as an effective inductive bias for sequence-based models lacking explicit spatial ordering, allowing them to learn more discriminative spatial representations. When the inter-slice scanning was removed, performance degradation became more pronounced, with the AUC decreasing to 78.16%. This highlights the crucial role of inter-slice scanning in preserving spatial continuity across 3D structures. It ensures that anatomically adjacent regions remain spatially coherent even after being unfolded into sequential form, thus facilitating the model's ability to leverage the inherent anatomical priors present in volumetric medical images.

Taken together, these results confirm the necessity of both multi-directional and inter-slice scanning in the I^2S mechanism. The synergistic effect of these two components enhances the model's capability to capture global dependencies and spatial coherence in 3D medical image analysis.

5 Conclusion

In this study, we propose MSCM-Net, a multi-scale CNN–Mamba hybrid framework for predicting pathological complete response (pCR) in lung cancer after neoadjuvant chemoimmunotherapy from CT scans. Its core innovation is a stage-wise parallel design that integrates a CNN branch and a Mamba branch with the proposed I^2S scanning mechanism. This architecture combines the CNN's inductive bias for local structural modeling with the Mamba's capacity to capture long-range dependencies, allowing the model to extract complementary local and global representations and achieve a multi-scale understanding of 3D tumor structures.

To further enhance cross-scale feature integration, we introduce a MAFF module that performs efficient channel-level and multi-scale spatial fusion between CNN and Mamba features. Experimental results on our lung cancer cohort demonstrate that MSCM-Net consistently outperforms state-of-the-art methods across multiple metrics, suggesting strong predictive capability and potential for clinical translation pending prospective, multi-center validation.

In summary, MSCM-Net provides an effective CT-based solution for pCR prediction in lung cancer and offers a generalizable hybrid design that may inform future research on treatment-response prediction across cancer types.

Acknowledgment

This work was supported in part by the Key Research projects of Hunan Education Department(No. 25A0545, No. 24A0575), the General Research Projects of Hunan Education Department (No. 25C0531), the National Natural Science Foundation of China (No. U24A20256, No. 62573434),

the Natural Science Foundation of Hunan Province (No. 2025JJ50374) and the High-Performance Computing Center of Central South University.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Reference

[1] W. D. Travis, S. Dacic, I. Wistuba, L. Sholl, P. Adusumilli, L. Bubendorf, P. Bunn, T. Cascone, J. Chaft, G. Chen *et al.*, Iaslc multidisciplinary recommendations for pathologic assessment of lung cancer resection specimens after neoadjuvant therapy, *Journal of Thoracic Oncology*, vol. 15, no. 5, pp. 709–740, 2020.

[2] Z. Jiang, Q. Li, J. Ruan, Y. Li, D. Zhang, Y. Xu, Y. Liao, X. Zhang, D. Gao, and Z. Li, Machine learning-based prediction of pathological responses and prognosis after neoadjuvant chemotherapy for non–small-cell lung cancer: A retrospective study, *Clinical Lung Cancer*, vol. 25, no. 5, pp. 468–478, 2024.

[3] M. Chen, S. J. Copley, P. Viola, H. Lu, and E. O. Aboagye, Radiomics and artificial intelligence for precision medicine in lung cancer treatment, in *Seminars in cancer biology*, vol. 93. Elsevier, 2023, pp. 97–113.

[4] T. P. Coroller, V. Agrawal, E. Huynh, V. Narayan, S. W. Lee, R. H. Mak, and H. J. Aerts, Radiomic-based pathological response prediction from primary tumors and lymph nodes in nsclc, *Journal of Thoracic Oncology*, vol. 12, no. 3, pp. 467–476, 2017.

[5] G. Ye, G. Wu, C. Zhang, M. Wang, H. Liu, E. Song, Y. Zhuang, K. Li, Y. Qi, and Y. Liao, Ct-based quantification of intratumoral heterogeneity for predicting pathologic complete response to neoadjuvant immunochemotherapy in non-small cell lung cancer, *Frontiers in Immunology*, vol. 15, p. 1414954, 2024.

[6] S. Fan, J. Xie, S. Zheng, J. Wang, B. Zhang, Z. Zhang, S. Wang, Y. Cui, J. Liu, X. Zheng, Z. Ye, X. Cui, and D. Yue, Non-invasive ct based multiregional radiomics for predicting pathologic complete response to preoperative neoadjuvant chemoimmunotherapy in non-small cell lung cancer, *European Journal of Radiology*, vol. 189, p. 112171, 2025.

[7] K. H. Leung, S. P. Rowe, M. S. Sadaghiani, J. P. Leal, E. Mena, P. L. Choyke, Y. Du, and M. G. Pomper, Deep semisupervised transfer learning for fully automated whole-body tumor quantification and prognosis of cancer on pet/ct, *Journal of Nuclear Medicine*, vol. 65, no. 4, pp. 643–650, 2024.

[8] J. Guo, B. Chen, H. Cao, Q. Dai, L. Qin, J. Zhang, Y. Zhang, H. Zhang, Y. Sui, T. Chen *et al.*, Cross-modal deep learning model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer, *NPJ Precision Oncology*, vol. 8, no. 1, p. 189, 2024.

[9] W. Qu, C. Chen, C. Cai, M. Gong, Q. Luo, Y. Song, M. Yang, and M. Shi, Non-invasive prediction for pathologic complete response to neoadjuvant chemoimmunotherapy in lung cancer using ct-based deep learning: a multicenter study, *Frontiers in Immunology*, vol. 15, p. 1327779, 2024.

[10] Y. She, B. He, F. Wang, Y. Zhong, T. Wang, Z. Liu, M. Yang, B. Yu, J. Deng, X. Sun *et al.*, Deep learning for predicting major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer: A multicentre study, *EBioMedicine*, vol. 86, 2022.

[11] A. N. Caragliano, F. Ruffini, C. Greco, E. Ippolito, M. Fiore, C. Tacconi, L. Nibid, G. Perrone, S. Ramella, P. Soda *et al.*, Doctor-in-the-loop: An explainable, multi-view deep learning framework for predicting pathological response in non-small cell lung cancer, *arXiv preprint arXiv:2502.17503*, 2025.

[12] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, Attention u-net: Learning where to look for the pancreas, *arXiv preprint arXiv:1804.03999*, 2018.

[13] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[14] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in *Proceedings of European Conference on Computer Vision*, 2022, pp. 205–218.

[15] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers, *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[16] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, U-net transformer: Self and cross attention for medical image segmentation, in *Proceedings of 12th International Workshop on Machine Learning in Medical Imaging*, 2021, pp. 267–276.

[17] H. Kuang, Y. Wang, J. Liu, J. Wang, Q. Cao, B. Hu, W. Qiu, and J. Wang, Hybrid cnn-transformer network with circular feature interaction for acute ischemic stroke lesion segmentation on non-contrast ct scans, *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2303–2316, 2024.

[18] A. Gu and T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, *arXiv preprint arXiv:2312.00752*, 2023.

[19] G. Li, Q. Huang, W. Wang, and L. Liu, Selective and multi-scale fusion mamba for medical image segmentation, *Expert Systems with Applications*, vol. 261, p. 125518, 2025.

[20] J. Liu, H. Yang, H.-Y. Zhou, L. Yu, Y. Liang, Y. Yu, S. Zhang, H. Zheng, and S. Wang, Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation, *IEEE Transactions on Medical Imaging*, 2024.

[21] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, Segmamba:

Long-range sequential modeling mamba for 3d medical image segmentation, in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024, pp. 578–588.

[22] A. Cao, Z. Li, J. Jomsky, A. F. Laine, and J. Guo, Medsegmamba: 3d cnn-mamba hybrid architecture for brain segmentation, *arXiv preprint arXiv:2409.08307*, 2024.

[23] H. Gong, L. Kang, Y. Wang, X. Wan, and H. Li, nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model, *arXiv preprint arXiv:2402.03526*, 2024.

[24] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, Medvit: a robust vision transformer for generalized medical image classification, *Computers in biology and medicine*, vol. 157, p. 106791, 2023.

[25] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, Repvit: Revisiting mobile cnn from vit perspective, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 909–15 920.

[26] Y. Shi, M. Sun, Y. Wang, J. Ma, and Z. Chen, Evit: An eagle vision transformer with bi-fovea self-attention, *IEEE Transactions on Cybernetics*, vol. 55, no. 3, pp. 1288–1300, 2025.

[27] X. Ma, X. Dai, Y. Bai, Y. Wang, and Y. Fu, Rewrite the stars, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[28] X. Ding, Y. Zhang, Y. Ge, S. Zhao, L. Song, X. Yue, and Y. Shan, Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5513–5524.

[29] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, Run, don't walk: chasing higher flops for faster neural networks, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 021–12 031.

[30] Y. Liu, Y. Su, J. Peng, W. Zhang, F. Zhao, Y. Li, X. Song, Z. Ma, W. Zhang, J. Ji *et al.*, A multidimensional deep ensemble learning model predicts pathological response and outcomes in esophageal squamous cell carcinoma treated with neoadjuvant chemoradiotherapy from pretreatment ct imaging: A multicenter study, *Radiotherapy and Oncology*, p. 111133, 2025.

[31] Z. Zhang, T. Luo, M. Yan, H. Shen, K. Tao, J. Zeng, J. Yuan, M. Fang, J. Zheng, I. Bermejo *et al.*, Voxel-level radiomics and deep learning for predicting pathologic complete response in esophageal squamous cell carcinoma after neoadjuvant immunotherapy and chemotherapy, *Journal for immunotherapy of cancer*, vol. 13, no. 3, p. e011149, 2025.

[32] H. Gong, L. Kang, Y. Wang, Y. Wang, X. Wan, X. Wu, and H. Li, nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model, in *Proceedings of 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 2025, pp. 1–5.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[34] R. Chang, S. Qi, Y. Zuo, Y. Yue, X. Zhang, Y. Guan, and W. Qian, Predicting chemotherapy response in non-small-cell lung cancer via computed tomography radiomic features: Peritumoral, intratumoral, or combined? *Frontiers in Oncology*, vol. 12, p. 915835, 2022.

[35] G. Ye, G. Wu, Y. Qi, K. Li, M. Wang, C. Zhang, F. Li, L. Wee, A. Dekker, C. Han *et al.*, Non-invasive multimodal ct deep learning biomarker to predict pathological complete response of non-small cell lung cancer following neoadjuvant immunochemotherapy: a multicenter study, *Journal for ImmunoTherapy of Cancer*, vol. 12, no. 9, p. e009348, 2024.

Jiancun Zhou received the master's degree in applied mathematics from Guizhou University, Guiyang, China, in 2006. He is currently a professor with Hunan City University, Yiyang, China. He is also pursuing the Ph.D. degree in computer science and engineering at Central South University, Changsha, China. His research interests include medical big data and artificial intelligence for medical data analysis.

Hulin Kuang received the Ph.D. degree in electronic engineering from the City University of Hong Kong, Hong Kong, China, in 2016. He was a postdoctoral researcher with the Department of Clinical Neurosciences, University of Calgary, Calgary, Canada, from 2017 to 2020. He is currently a distinguished associate professor with the School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include medical image analysis and deep learning.

Jianxin Wang received the bachelor's and master's degrees in computer engineering from Central South University, Changsha, China, in 1992 and 1996, respectively, and the Ph.D. degree in computer science from Central South University, Changsha, China, in 2001. He is currently the dean and a professor with the School of Computer Science and Engineering, Central South University, Changsha, China. His current research interests include algorithm analysis and optimization, parameterized algorithms, bioinformatics, and computer networks.