


## Article

# Medical Image Classification with a Hybrid SSM Model Based on CNN and Transformer

Can Hu <sup>1</sup>, Ning Cao <sup>1,\*</sup>, Han Zhou <sup>2</sup> and Bin Guo <sup>3</sup><sup>1</sup> School of Computer and Soft, Hohai University, Nanjing 211100, China; hucan@hhu.edu.cn<sup>2</sup> School of Electronic Science and Engineering, Nanjing University, Nanjing 210046, China<sup>3</sup> College of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China

\* Correspondence: caoning@hhu.edu.cn

**Abstract:** Medical image classification, a pivotal task for diagnostic accuracy, poses unique challenges due to the intricate and variable nature of medical images compared to their natural counterparts. While Convolutional Neural Networks (CNNs) and Transformers are prevalent in this domain, each architecture has its drawbacks. CNNs, despite their strength in local feature extraction, fall short in capturing global context, whereas Transformers excel at global information but can overlook fine-grained details. The integration of CNNs and Transformers in a hybrid model aims to bridge this gap by enabling simultaneous local and global feature extraction. However, this approach remains constrained in its capacity to model long-range dependencies, thereby hindering the efficient extraction of distant features. To address these issues, we introduce the MambaConvT model, which employs a state-space approach. It begins by locally processing input features through multi-core convolution, enhancing the extraction of deep, discriminative local details. Next, depth-separable convolution with a 2D selective scanning module (SS2D) is employed to maintain a global receptive field and establish long-distance connections, capturing the fine-grained features. The model then combines hybrid features for comprehensive feature extraction, followed by global feature modeling to emphasize on global detail information and optimize feature representation. This paper conducts thorough performance experiments on different algorithms across four publicly available datasets and two private datasets. The results demonstrate that MambaConvT outperforms the latest classification algorithms in terms of accuracy, precision, recall, F1 score, and AUC value ratings, achieving superior performance in the precise classification of medical images.

**Keywords:** medical images; image classification; state-space models; transformers; convolutional neural network



**Citation:** Hu, C.; Cao, N.; Zhou, H.; Guo, B. Medical Image Classification with a Hybrid SSM Model Based on CNN and Transformer. *Electronics* **2024**, *13*, 3094. <https://doi.org/10.3390/electronics13153094>

Academic Editor: Luca Mesin

Received: 21 June 2024

Revised: 29 July 2024

Accepted: 31 July 2024

Published: 5 August 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The evolution of modern medical research is intrinsically linked to the pivotal role of medical images, which facilitate the visualization of physiological structures and disease-related information. These images are instrumental in aiding medical professionals and scientists in diagnosing patients and devising targeted treatment plans [1]. Over the past few decades, a plethora of imaging techniques have emerged, including computed tomography (CT) [2], ultrasound, X-ray [3], endoscopic imaging, and magnetic resonance imaging (MRI) [4,5], playing a crucial role in early disease detection and management. However, the surge in medical image data has exposed significant challenges in traditional manual analysis [6,7]. Clinicians' varying expertise can lead to inconsistent disease detection and diagnosis, and the visual nuances of medical images can be difficult to discern, with subtle variations sometimes posing a challenge. Moreover, medical images are often voluminous, high-dimensional, and lack comprehensive labeling and may suffer from quality issues. To address these limitations, Computer-Aided Diagnosis (CAD) [8] systems have been developed to enhance clinicians' image interpretation efficiency and promote more accurate, objective diagnostic decisions.

Traditional CAD systems offer objective support to medical and scientific researchers by extracting, selecting, and classifying features from medical images. Among these steps, feature extraction holds paramount importance. Although various techniques exist, such as filter-based features [9,10], scale-invariant feature transforms [11], and local binary patterns [12,13], their manual setup and extensive task-specific tuning render traditional CAD systems expensive and less efficient for medical imaging applications. Deep learning, leveraging its ability to extract abstract features directly from raw image data, demonstrates significant potential in medical image processing and is widely employed in tasks like image classification, segmentation, and target detection [14–16]. As a fundamental task in clinical diagnosis, medical image classification stands out among visual tasks, playing a pivotal role [1,17].

Numerous studies have demonstrated that models based on Convolutional Neural Networks (CNNs) and Transformers achieve superior performance in medical image classification tasks. Although CNNs are adept at extracting local features [18–20], their models are constrained by the local receptive field, which significantly impedes their capacity to capture global contextual information and establish long-range dependencies, thereby leading to insufficient feature extraction and unsatisfactory classification outcomes. Initially conceived for natural language processing (NLP), Vision Transformers have been extensively adopted in computer vision tasks. Vision Transformers represent input images as a sequence of image patches and employ cascaded self-attention modules to effectively extract global contextual features, yet they suffer from notable deficiencies in local feature extraction. Recognizing the complementarity between CNNs and Vision Transformers, researchers have proposed a series of hybrid architectures [21,22] that integrate CNNs and Vision Transformers to more effectively address a variety of medical image challenges. Despite their exceptional performance in extracting local–global features, these hybrid models still exhibit limitations in remote modeling, failing to effectively extract features at a distance, which in turn leads to the neglect of critical distant features and adversely affects their performance in medical image dense detection tasks. Considering the significance of medical image classification tasks and the limitations of current models, the exploration of an efficient medical image classification model holds great importance for the advancement of intelligent clinical diagnosis and computer vision.

In recent years, State-Space Models (SSMs) [23,24] have garnered substantial attention from researchers due to their ability to establish long-range dependencies and exhibit linear complexity with respect to input size. SSM-based models have been extensively investigated in various domains, including natural language understanding [23,25], image segmentation [26], and image classification [27]. Inspired by VMamba's [28] success in natural image classification, this study presents an SSM-based MambaConvT model, with a core component called Conv-SSM-T. The model is designed to showcase its potential in medical image classification tasks, where images often exhibit high similarity, necessitating efficient extraction of both fine and coarse features. Conv-SSM-T integrates convolutional layers' local feature extraction prowess, SSMs' capacity for long-distance dependencies, and Transformer's global modeling capabilities, to effectively capture and accurately classify medical image features.

The main contributions of this paper can be summarized as follows:

- An SSM-based MambaConvT model is proposed and its potential application in medical image classification tasks is explored.
- Full experiments were conducted on six datasets, four of which are public and two are private, and the results show that the MambaConvT model has excellent performance on different classification tasks.
- It provides a valuable reference for the future development of more effective SSM-based medical classification models.

The rest of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the overall MambaConvT framework and detailed information about its core module Conv-SSM-T, Section 4 shows the detailed experimental procedure of MambaConvT on different medical classification tasks and its results, and Section 5 draws the experimental conclusions of this paper.

## 2. Related Works

### 2.1. CNN Module

In recent years, CNNs have made significant contributions to the field of computer vision due to their unique ability to extract deep discriminative features. Howard et al. [29] proposed the MobileNets model, which uses depthwise separable convolutions to design a lightweight CNN network, improving the efficiency of CNNs. Zhang et al. [30] proposed the ShuffleNet model, which uses pointwise group convolutions and channel shuffling to maintain model accuracy while significantly reducing computational costs. Koonce et al. [31] proposed the ResNet network structure, which uses batch normalization to accelerate training, effectively alleviating the problem of gradient vanishing and ensuring that deeper networks capture image details and features. Zhu et al. [32] adopted the ResNet network concept and proposed a more aggressive dense connection mechanism, achieving better performance than ResNet by reusing features through channel connections. Woo et al. [33] proposed the ConvNext model, which first introduced depthwise separable convolutions to reduce the number of model parameters and improve computational efficiency; it also used larger convolution kernels and strides to cover a wider range of input features while reducing information loss. Experiments showed that ConvNext can effectively capture details and contextual information in images, improving the performance of the model on complex visual tasks. In the field of precise classification of medical images, Hosein et al. [18] proposed a C-Net architecture based on CNN, which connects multiple CNN networks as feature extractors, inputting them into an internal network for classification, achieving excellent results in the classification of histopathological images. Kumar et al. [19] proposed a CNN-based transfer learning strategy model to detect and classify brain images, preprocessing the images and using binarization methods to improve image quality, performing exceptionally well in the classification of normal, benign, and malignant brain tumors. Ashwath et al. [34] proposed the TS-CNN network architecture, which has three branches: a global branch, an attention branch, and a fusion branch that extracts knowledge from the global and local branches for classification, demonstrating excellent accuracy in classification tasks while having inherent interpretability.

### 2.2. Transformer Module

Han et al. [35] introduced the Transformer model, which has demonstrated significant success in NLP. Inspired by this, an increasing number of studies have applied the Transformer model to visual tasks, including image classification, semantic segmentation, and object detection. Han et al. [36] proposed the Vision Transformer, which diverged from the Transformer's use of position encoding to capture relative word positions in the input sequence. Instead, the input image was divided into a series of image blocks, each flattened into a vector and then represented through position encoding and embedding vectors. A multi-head self-attention mechanism was introduced to handle the relationships between image blocks. Experiments demonstrated the Vision Transformer's exceptional performance in computer vision tasks. Yuan et al. [37] introduced the T2T-ViT model to address the limitations of the Vision Transformer. In each token-to-token step, the model reconstructed the output tokens into an image, then used soft segmentation to tile and aggregate the surrounding tokens into new tokens, embedding the local structure into the generated tokens and inputting them to the next Transformer layer. T2T-ViT adopted a deep and narrow structure, significantly enhancing feature richness. Liu et al. [38] proposed the Swin Transformer, designed to handle large variance in visual entities, high image resolution, and numerous pixels. The model introduced a new sliding window mechanism, limiting attention calculation within a window to introduce CNN convolution locality while saving computational resources. It also used a hierarchical downsampling design to gradually increase the field of view, allowing the attention mechanism to notice global features. In medical image classification, Manzari et al. [21] proposed MedViT, a model combining CNN's local feature capture and Vision Transformer's global connectivity, demonstrating outstanding performance on the MedMNIST-2D dataset. Wu et al. [22] proposed CTransCNN, featuring three main components in both CNN and

Transformer branches: a multi-label enhanced feature module with multi-head attention, a multi-branch residual module, and an information interaction module. The model excelled in multi-label medical image classification tasks.

### 2.3. State-Space Module

Inspired by the continuous state-space model in controlled systems and combining the HiPPO initialization method proposed by Gu et al. [39], the LSSL proposed by Gu et al. [24] has demonstrated potential in handling long-range dependency issues. However, LSSL requires high computational and memory demands, making it challenging to apply in reality. To address these issues, Gu et al. [23] proposed S4, which normalizes parameters into a diagonal structure. In addition to S4, models such as complex diagonal structures, multi-input–multi-output support, and selection mechanisms have been proposed. These models primarily focus on the application of state space in remote and arbitrary data, such as language and speech, for tasks like language understanding and content reasoning, with little attention to visual tasks. The S4ND proposed by Nguyen et al. [40] is the first to apply state space mechanisms to visual tasks, and it has shown potential to compete with Vision Transformers. However, S4ND simply extended the S4 model and failed to effectively capture image information in a dependency-aware manner. Building on S4ND, Gu et al. [25] proposed Mamba, which introduced a selective scanning mechanism capable of efficiently capturing image information, with great potential in the field of computer vision. In the medical imaging field, Ma et al. [26] integrated Mamba into UNet to enhance the modeling of remote dependencies in CNNs and proposed a universal network for medical image segmentation that is applicable to both three-dimensional and two-dimensional images. Xing et al. [41] combined the U-shaped structure with Mamba to propose the Segmamba model, which models global features holistically across different scales. Ma et al. [42] constructed a medical image segmentation model based on the SSM model, establishing a new baseline for medical image segmentation.

Based on the reported deficiencies of CNNs and Transformers in information capture and the advantage of Mamba in establishing long-range dependencies, this paper proposes the MambaConvT model, which is based on the SSM. This model combines the local feature extraction ability of convolutional layers, the ability to establish long-range dependencies through the SSM, and the global modeling ability of Transformers, aiming to effectively extract and accurately classify medical image features, thereby further promoting intelligent clinical diagnosis and the development of computer vision.

## 3. Methodology

### 3.1. Mamba Module

The SSM-based Mamba model relies on a classical continuous system that maps a one-dimensional input function or sequence  $x(t) \in \mathcal{R}$  to an output  $y(t) \in \mathcal{R}$  through an intermediate implicit state  $h(t) \in \mathcal{R}^N$  [28]. The above process can be expressed as a linear ordinary differential equation:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \quad (1)$$

$$y(t) = \mathbf{C}h(t) \quad (2)$$

where  $\mathbf{A} \in \mathcal{R}^{N \times N}$  denotes the state matrix, while  $\mathbf{B} \in \mathcal{R}^{N \times 1}$  and  $\mathbf{C} \in \mathcal{R}^{N \times 1}$  denote the projection parameters.

The Mamba model discretizes the classical continuous system in order to better integrate with deep learning scenarios by introducing the time scale parameter  $\Delta$  and using a zero-order hold discretization rule to transform  $\mathbf{A}$  and  $\mathbf{B}$  into the discrete parameters  $\overline{\mathbf{A}}$  and  $\overline{\mathbf{B}}$ ; the above process can be expressed as follows [28]:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}) \quad (3)$$

$$\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \quad (4)$$

After the above discretization operation, the SSM-based Mamba model can be computed in two ways, linear recursion or global convolution, defined as follows:

$$\begin{aligned} h'(t) &= \bar{\mathbf{A}}h(t) + \bar{\mathbf{B}}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \quad (5)$$

$$\begin{aligned} \bar{\mathbf{K}} &= (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \dots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}}) \\ y &= x * \bar{\mathbf{K}} \end{aligned} \quad (6)$$

where  $\bar{\mathbf{K}} \in \mathcal{R}^L$  represents a structured convolutional kernel, and  $L$  denotes the length of the input sequence  $x$ .

### 3.2. MambaConvT Architecture

The overall network architecture of MambaConvT is shown in Figure 1a. Specifically, the MambaConvT network architecture includes a patch embedding layer, a Conv-SSM-T module, and a patch merging layer. The patch embedding layer firstly divides the input medical image  $x \in \mathcal{R}^{H \times W \times 3}$  into image patches with a size of  $4 \times 4$  but instead of further spreading the patches into one-dimensional sequences as in the Vision Transformer model, the 2D structure of the image is preserved. The dimensions of the image are then mapped to a specified value  $C$  to obtain the embedded image  $x' \in \mathcal{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ . Before inputting  $x'$  into the backbone network, it needs to be normalized, which is divided into four steps, and after the first three steps both the height and width of the image features can be halved while doubling the feature dimension number  $C$ .

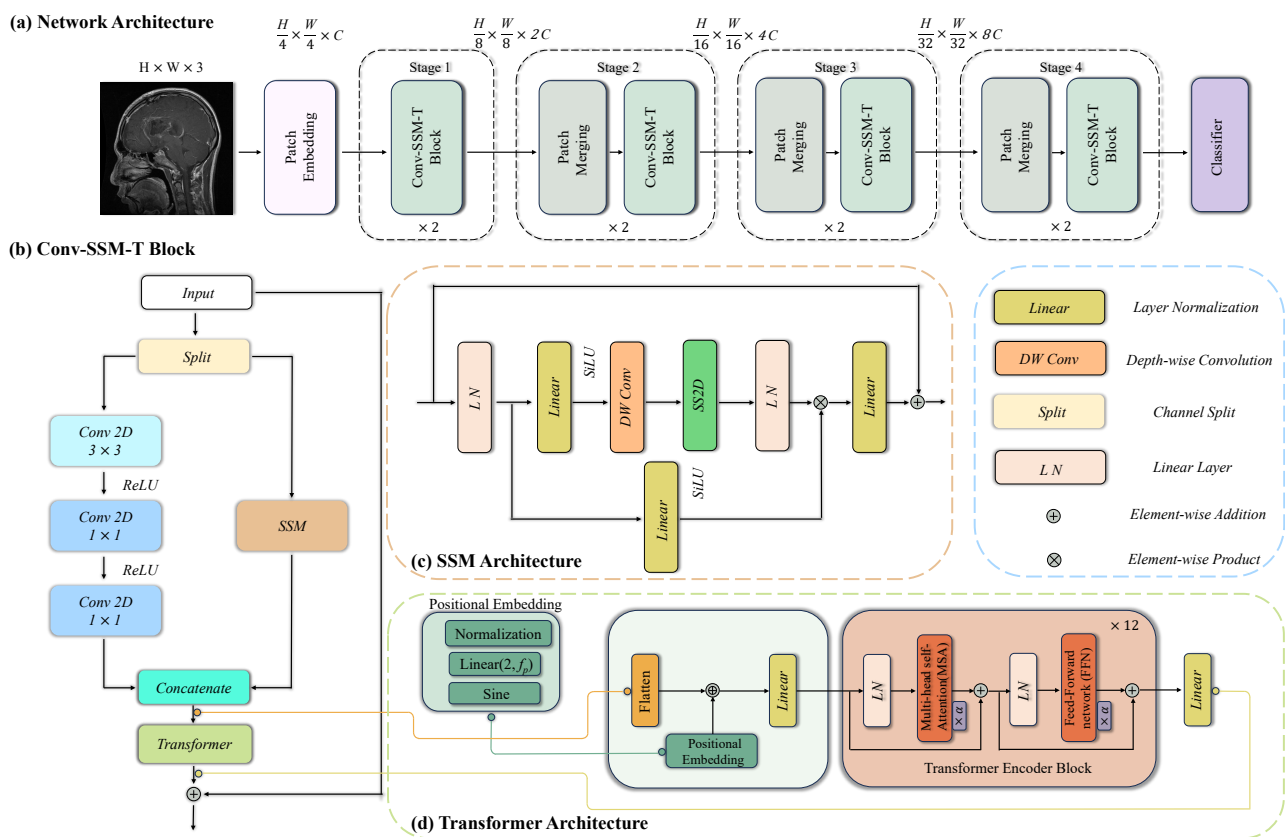


Figure 1. Overall framework of MambaConvT.



### 3.3. Conv-SSM-T Module

The Conv-SSM-T module is the core module of the MambaConvT architecture, as shown in Figure 1b. Specifically, Conv-SSM-T, as a dual-branch module, first utilizes channel partitioning to split the input into two equal-sized sub-inputs, one of which is a convolutional branch and the other is an SSM branch, as shown in Figure 1c. In the convolutional branch, this paper adopts a simple continuous convolution to model local features on its input and uses the ReLU activation function, which aims to focus on the local details of the image and extract deep discriminative features. In the SSM branch, this paper first performs a normalization operation, which is divided into two sub-branches after layer normalization. One of them passes through the linear layer and activation function SiLU; the input of the other branch passes through the linear layer, depth-separable convolution, and activation function SiLU before further feature extraction through the 2D selective scanning module SS2D, aiming at preserving the global receptive field and establishing the long-distance dependency to capture the distal detail features and then multiply the two sub-branches one by one and merge them together as the SSM branch's output. Finally, the convolutional branch and SSM branch are feature-mixed and passed into the Transformer model for global feature modeling, aiming at focusing more global detail information and extracting fuller image features. The Transformer model is shown in Figure 1d, specifically, it mainly consists of a stack of coding blocks with input dimensions  $(w, h, f)$ . The model first unfolds the encoding along the spatial dimension and then connects with a 2D Fourier positional encoding embedding of dimension  $f_p$ . In order to improve convergence, this paper employs a Rezero regularization scheme and introduces a trainable scale parameter  $\alpha$  to modulate the magnitude of the nontrivial branches of the residual block, and finally the result is linearly mapped as the output of the Conv-SSM-T module.

### 3.4. SS2D Module

The SS2D module within the SSM branch inherits the selective scanning space-state sequence model S6 from the NLP domain, addressing its directional sensitivity issues. To bridge the gap between one-dimensional array scanning and two-dimensional plane traversal, SS2D introduces a cross-scan module (CSM). This module employs a four-directional scanning strategy, traversing the spatial domain of the entire feature map from the four corners to the opposite positions, ensuring that each pixel integrates information from all other locations in different directions without increasing the linear computational complexity, while receiving global information and capturing long-range dependencies simultaneously.

SS2D comprises three components: scanning expansion, S6 block, and scanning merge. Figure 2 provides an intuitive illustration of the SS2D's internal mechanism. The scanning expansion operation initially expands the input image into a sequence by scanning in four different directions (left up to right down, right down to left up, right up to left down, and left down to right up). Subsequently, the S6 block processes all sequences to extract features, ensuring a comprehensive scan of information from all directions. Ultimately, the four outputs from the four directions are integrated through scanning merge to construct the final two-dimensional feature map, resulting in an output with the same size as the input. The S6 block builds upon S4 by introducing a selection mechanism, adjusting the parameters of the input SSM to allow the model to distinguish and retain relevant information while filtering out irrelevant information. Algorithm 1 offers the pseudocode for the S6 block.

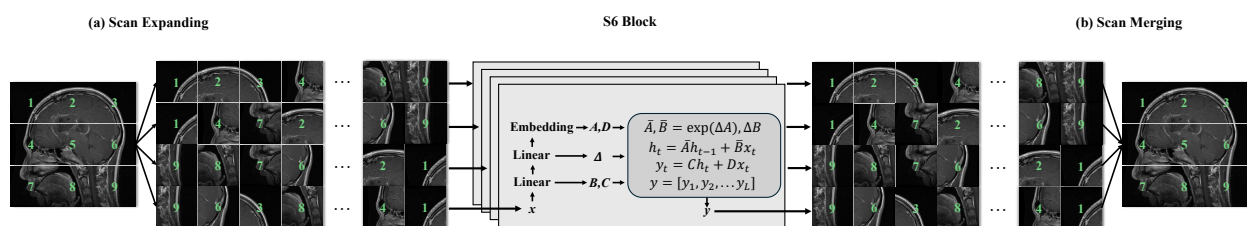


Figure 2. Visual display of the internal modeling process of SS2D.

**Algorithm 1:** Pseudo-code for S6 block in SS2D

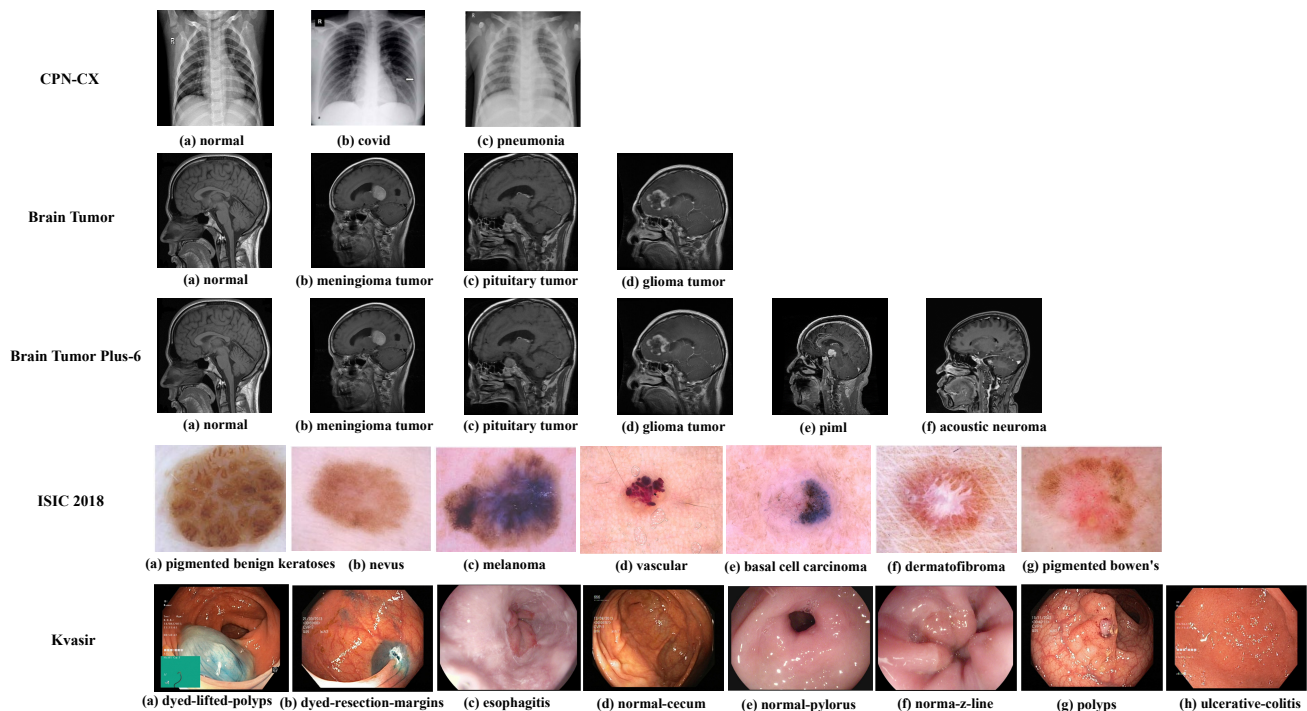
---

**Input :**  $x$ , the feature with shape  $[B, L, D]$  (batch size, token length, dimension)  
**Params :**  $A$ , the nn.Parameter;  $D$ , the nn.Parameter  
**Operator :**  $\text{Linear}(\cdot)$ , the linear projection layer  
**Output :**  $y$ , the feature with shape  $[B, L, D]$   
1:  $\Delta, B, C = \text{Linear}(x), \text{Linear}(x), \text{Linear}(x)$   
2:  $\bar{A} = \exp(\Delta A)$   
3:  $\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B$   
4:  $h_t = \bar{A}h_{t-1} + \bar{B}x_t$   
5:  $y_t = Ch_t + Dx_t$   
6:  $y = [y_1, y_2, \dots, y_t, \dots, y_L]$   
7: **return**  $y$

---

**4. Experiments and Results****4.1. Datasets**

ISIC 2018 dataset [43]: this dataset was released by the International Skin Imaging Collaborative (ISIC) and contains a total of 10015 images in seven different categories. These include dermatofibromas (115), vascular lesions (142), actinic keratoses (327), basal cell carcinomas (514), benign keratoses (1099), melanomas (1113), and melanocytic naevi (6705), and all the images have sizes of  $650 \times 450$  pixels, as exemplified by row 4 of Figure 3.



**Figure 3.** Example of a dataset demonstration.

Kvasir dataset [44]: this dataset was collected using the Norwegian Vestre Viken endoscopic device and was annotated and validated by several experienced endoscopists. The dataset was divided into eight different categories, 500 sheets each, which included pathologies related to the gastrointestinal tract (esophagitis, polyps, ulcerative colitis), anatomical landmarks (normal cecum, normal pylorus, normal z-line), and two other categories (stained resection margins, stained polyps), and the dataset's resolution ranged from  $720 \times 576$  to  $1920 \times 1072$ , as exemplified by row 5 of Figure 3.

COVID-19-Pneumonia-Normal Chest X-ray Images (CPN-CX) dataset: the data were divided into three categories, namely, 1802 normal chest radiographs, 1626 chest radio-

graphs of new coronary pneumonia, and 1800 chest radiographs of viral pneumonia. The new crown pneumonia chest radiographs were collected from several public sources such as Github, German Medical School, and SIRM, while the normal chest radiographs and viral pneumonia chest radiographs were obtained from Kaggle’s “Chest X-ray Images (pneumonia)” database. All of its images were resized to  $256 \times 256$  pixels, as shown in row 1 of Figure 3.

Brain Tumor dataset [45]: this dataset is from Kaggle’s Brain Tumor Classification database, with a total of 3264 MRI images, including 4 different categories, including 926 gliomas, 901 pituitary tumors, 937 meningiomas, and 500 non-tumors. The resolution of the datasets were all adjusted to  $256 \times 256$ , and an example is shown in row 2 of Figure 3.

Brain Tumor Plus dataset: the non-tumor images in the Brain Tumor dataset were identified by the radiologists of the Jiangsu Provincial People’s Hospital as not being in the same MRI scanning sequence as the other three categories, which may lead to an impact on the classification effect. The Brain Tumor Plus dataset is based on the Brain Tumor dataset. Additional non-tumor images from the Radiology Department of Jiangsu Provincial People’s Hospital were imported from the same sequence as the other three categories, and the data volume was doubled, i.e., 1852 neurogliomas, 1802 pituitary tumors, 1874 meningiomas, and 1000 non-tumors. The dataset resolutions were all adjusted to  $256 \times 256$ .

Brain Tumor Plus-6 dataset: this dataset is based on the Brain Tumor Plus dataset, with additional images of acoustic neuromas and lymphomas imported from the radiology department of Jiangsu Provincial People’s Hospital in the same sequence as the other four categories. Among them, there were 1850 images of auditory neuroma and 1850 images of lymphoma. The resolution of the datasets were all adjusted to  $256 \times 256$ , and examples are shown in row 3 of Figure 3.

The datasets selected for this paper, including ISIC 2018, Kvasir, CPN-CX, and Brain Tumor, have been meticulously identified and annotated by clinical experts. These datasets have been recognized as crucial in various competitions, such as the MICCAI challenge and the Kaggle Brain Tumor Classification Challenge. They are also highly regarded by researchers and are widely utilized as authoritative sources in different medical image classification tasks. The private datasets Brain Tumor Plus and Brain Tumor Plus-6 enhance the Brain Tumor dataset, having been professionally identified and annotated by clinical doctors, ensuring authenticity and reliability. We confirmed that all methods were carried out in accordance with relevant guidelines and regulations, and informed consent for patients was waived by the Research Ethics Committee of the Nanjing Medical University. All experimental protocols and data in this study were approved by the Research Ethics Committee of the Nanjing Medical University. Approval number: NMUE2021301.

#### 4.2. Evaluation Metrics

In this paper, we use accuracy, precision, recall and F1 score as the evaluation indexes of the model according to the characteristics of medical images. The high and low values of these metrics can reflect the performance of the model, and their metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$



where  $TP$ ,  $FP$ ,  $TN$  and  $FN$  denote the number of true positive cases, false positive cases, true negative cases and false negative cases respectively. The values of all the metrics are in  $[0, 1]$ , where a larger value means a better performance of the assessment model.

#### 4.3. Compare Models

In order to demonstrate the effectiveness of the model, several representative classes of benchmark algorithms are selected for comparison in this paper. These include the CNN-based algorithms ResNet [31], ConvNeXt [33], and VGG-19 [46] and the Transformer-based algorithms ViT-B [36] and CNN-Transformer-based hybrid algorithms Swin-T [38].

#### 4.4. Implementation Details

Before training the network, we resize all images to  $224 \times 224 \times 3$  and normalize each image. During training, this paper uses the Adam optimizer and adopts a cross-entropy loss to optimize the model parameters. The specific hyperparameters of the model are shown in Table 1. All algorithms were tested on a Linux system equipped with four NVIDIA Tesla V100 for our experiments. The implementations used Python 3.6 (available at <https://www.python.org/downloads/release/python-3615/> accessed on 28 July 2024) and PyTorch 1.13.0 (<https://download.pytorch.org/whl/cu117> accessed on 28 July 2024). To ensure the fairness of the experiments, all algorithms were trained for 500 rounds, the batch size was set to 64, and an early abort strategy was used to prevent model overfitting.

**Table 1.** Hyperparameter settings for the model.

Hyperparameter	Value	Remark
Learning rate	0.001	Controls the step size of each parameter update
$\beta_1$	0.9	First-order exponential decay factor used to calculate the gradient
$\beta_2$	0.999	Two-order exponential decay factor used to calculate the gradient
Weight decay	$1 \times 10^{-4}$	Control the amplitude of the parameters to prevent overfitting

#### 4.5. Quantitative Results Analysis

As shown in Tables 2 and 3, on the simpler three-classification CPN-CX dataset, this paper's method improves 2.48%, 1.51%, 1.61%, and 1.39% in accuracy, F1 score, precision rate, and recall rate compared to the top-performing Swin-T algorithm in the benchmark. On the four-classification Brain Tumor dataset, this paper's method garners improvements of 2.25%, 1.45%, 1.25%, 1.64% in accuracy, F1 score, precision rate, and recall rate, respectively, compared to the best-performing Swin-T algorithm benchmark. Also, as shown by the error bars in Figures 4 and 5, the error margins demonstrated by this paper's method are the smallest on both datasets, highlighting the robustness of the method. As shown in Table 4, On the more intricate seven-class ISIC 2018 dataset, our method registers enhancements of 1.98% in accuracy, 1.29% in F1 score, 1.00% in precision, and a 0.34% boost in recall compared to the best Swin-T variant. The error range demonstrated by the method in this paper is also minimized as shown by the error bars in Figure 6. As shown in Table 5, On the challenging eight-class Kvasir dataset, our approach maintains strong results, with gains of 1.88% in accuracy, 1.22% in F1 score, 1.31% in precision, and 1.13% in recall compared to VGG-19's peak performance. The error range demonstrated by the method in this paper is also minimized as shown by the error bars in Figure 7. Consistently, our proposed method demonstrates superior performance when applied to private image classification datasets, as shown in Tables 6 and 7, as evidenced by the improvements of 2.11% in accuracy, 1.46% in F1 score, 0.90% in precision, and 2.04% in recall on the four-class Brain Tumor Plus, and 2.06% in accuracy, 1.34% in F1 score, 2.43% in precision, and 1.25% in recall on the six-class Brain Tumor Plus-6, compared to the best Swin-T performance. As shown by the error bars in Figures 8 and 9, the error margins demonstrated by this paper's method are likewise minimal, validating the robustness and robustness of this paper's method.

**Table 2.** Performance comparison of different algorithms for three classification tasks on CPN-CX dataset.

Methods	Acc(%) ↑	FI(%) ↑	Prec(%) ↑	Recall(%) ↑
ResNet	93.46	93.51	93.78	93.25
ConvNeXt	94.06	93.94	94.09	93.79
VGG19	94.12	94.05	94.22	93.89
ViTB	94.36	94.39	94.53	94.26
Swin-T	95.02	95.61	95.44	95.79
Ours	97.38	97.05	96.98	97.12

**Table 3.** Performance comparison of different algorithms for four classification tasks on Brain Tumor dataset.

Methods	Acc(%) ↑	FI(%) ↑	Prec(%) ↑	Recall(%) ↑
ResNet	87.21	86.32	87.14	85.52
ConvNeXt	86.27	85.88	86.74	85.03
VGG-19	84.33	85.03	85.15	84.91
ViT-B	88.15	88.36	88.75	87.98
Swin-T	89.02	89.82	89.34	90.31
Ours	91.02	91.12	90.46	91.79

**Table 4.** Performance comparison of different algorithms for seven classification tasks on ISIC 2018 dataset.

Methods	Acc(%) ↑	FI(%) ↑	Prec(%) ↑	Recall(%) ↑
ResNet	79.76	78.06	78.03	78.10
ConvNeXt	77.34	76.01	76.14	75.88
VGG-19	78.76	76.37	76.49	76.26
ViT-B	79.64	77.53	77.86	77.20
Swin-T	80.41	79.30	79.36	79.24
Ours	82.00	80.32	80.15	79.51

The proposed method's key strength lies in MambaConvT's formidable feature extraction prowess. As a component of MambaConvT, the CNN employs consecutive multi-kernel convolutions to locally analyze input features, with a focus on discerning local nuances in medical imagery and extracting deep, discriminative characteristics. To optimize computational efficiency, it employs parameter sharing. The SSM, another branch of MambaConvT, integrates depth-wise separable convolutions alongside the SS2D, enabling it to preserve a global receptive field and establish long-range dependencies to capture fine-grained details at a distance. Lastly, the Transformer concludes MambaConvT's sequence by globally modeling the concatenated features, thereby enhancing the detection of global details and enriching the extracted image features.

**Table 5.** Performance comparison of different algorithms for eight classification tasks on Kvasir dataset.

Methods	Acc(%) ↑	FI(%) ↑	Prec(%) ↑	Recall(%) ↑
ResNet	76.90	76.54	76.91	76.18
ConvSeXt	74.54	74.63	74.62	74.63
VGG-19	77.75	77.78	77.82	77.74
ViT-B	76.03	76.09	76.23	75.96
Swin-T	77.32	77.25	77.24	77.26
Ours	79.21	78.73	78.84	78.62

**Table 6.** Performance comparison of different algorithms for four classification tasks on Brain Tumor Plus dataset.

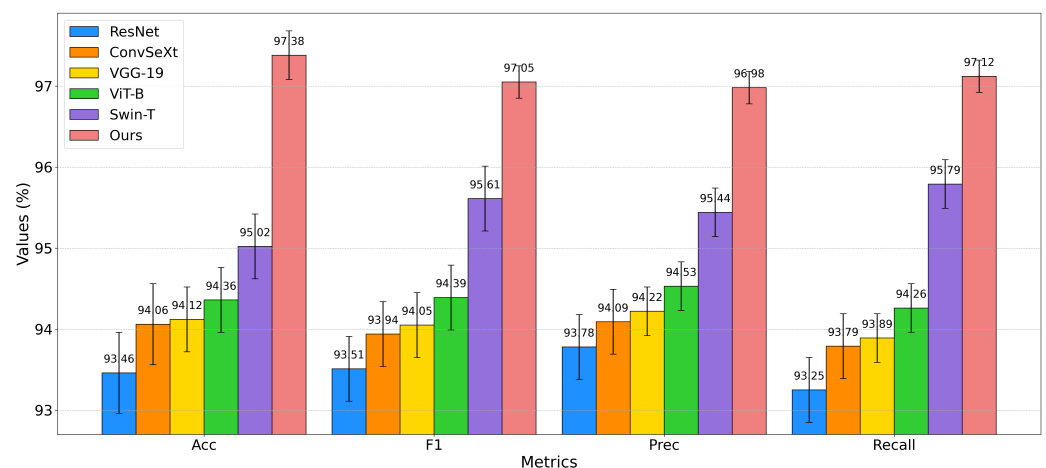
Methods	Acc(%) ↑	F1(%) ↑	Prec(%) ↑	Recall(%) ↑
ResNef	89.99	88.74	89.03	88.45
ConvNeXt	88.14	88.43	88.48	88.39
VGG-19	86.18	86.05	86.31	85.79
ViT-B	90.18	89.91	89.69	90.13
Swin-T	91.33	91.03	90.48	91.59
Ours	93.26	92.36	91.29	93.46

**Table 7.** Performance comparison of different algorithms for six classification tasks on Brain Tumor Plus-6 dataset.

Methods	Acc(%) ↑	F1(%) ↑	Prec(%) ↑	Recall(%) ↑
ResNet	87.01	86.50	86.23	86.77
ConvNeXt	86.03	85.28	85.14	85.42
VGG-19	83.11	83.76	83.64	83.88
ViT-B	88.32	88.39	88.02	88.77
Swin-T	88.56	89.10	88.87	89.33
Ours	90.38	90.29	90.13	90.45

A comparison of Tables 3 and 6 reveals that standardized data format and abundant data volume contribute to enhanced performance in medical image classification. Specifically, our method exhibits improvements of 2.46% in accuracy, 1.36% in F1 score, 0.92% in precision, and 1.82% in recall when applied to the four-class Brain Tumor Plus dataset compared to the four-class Brain Tumor dataset.

To thoroughly evaluate the proposed method's application performance against established benchmarks, this study opted for ViT-B, Swin-T, and the proposed method, focusing on datasets with high precision: CPN-CX and Brain Tumor Plus. ROC curves, a standard evaluation tool for binary classification models, were employed to analyze the results. As illustrated in Figure 10, the average AUC for our method stands at 0.976 on CPN-CX and 0.950 on Brain Tumor Plus, outperforming both ViT-B and Swin-T. The curve positioned in the top-left corner is indicative of superior performance, with a larger AUC area, suggesting that the proposed method exhibits greater potential for image classification tasks.

**Figure 4.** Visualization of performance between different algorithms on CPN-CX dataset.

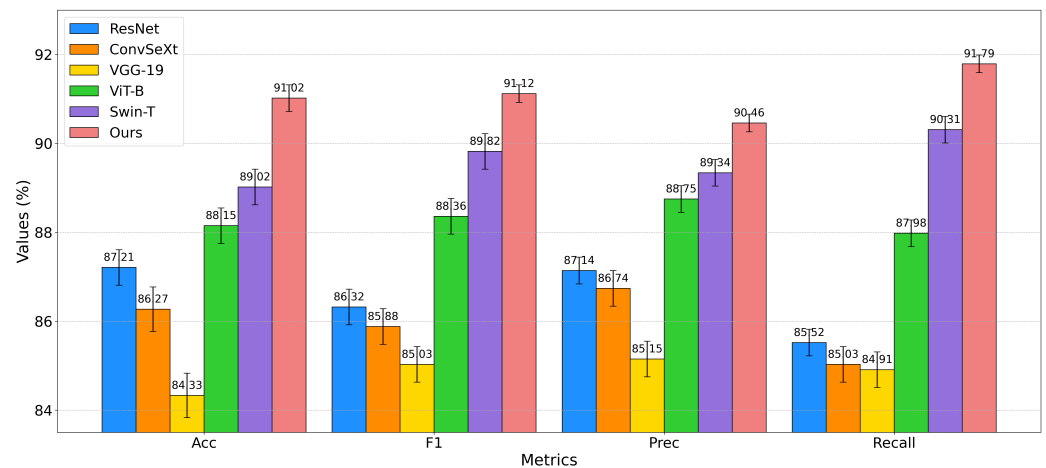


Figure 5. Visualization of performance between different algorithms on Brain Tumor dataset.

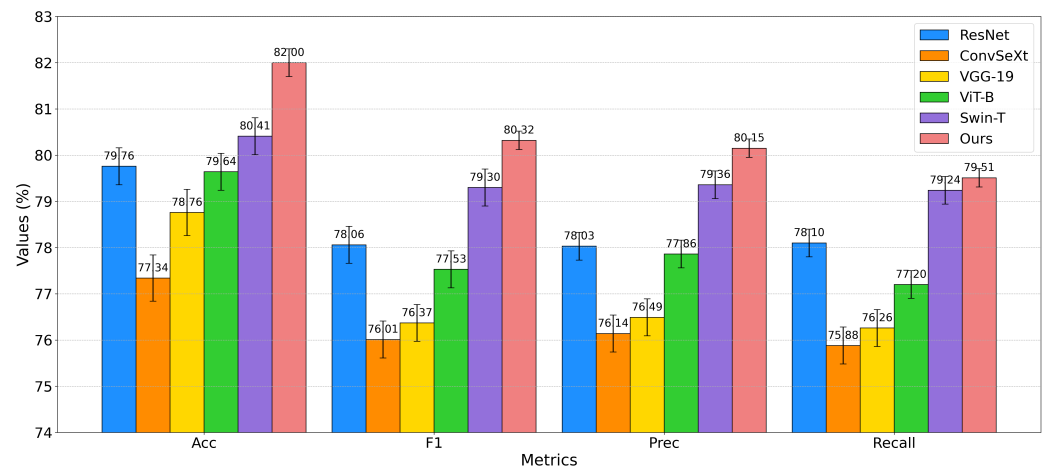


Figure 6. Visualization of performance between different algorithms on ISIC 2018 dataset.

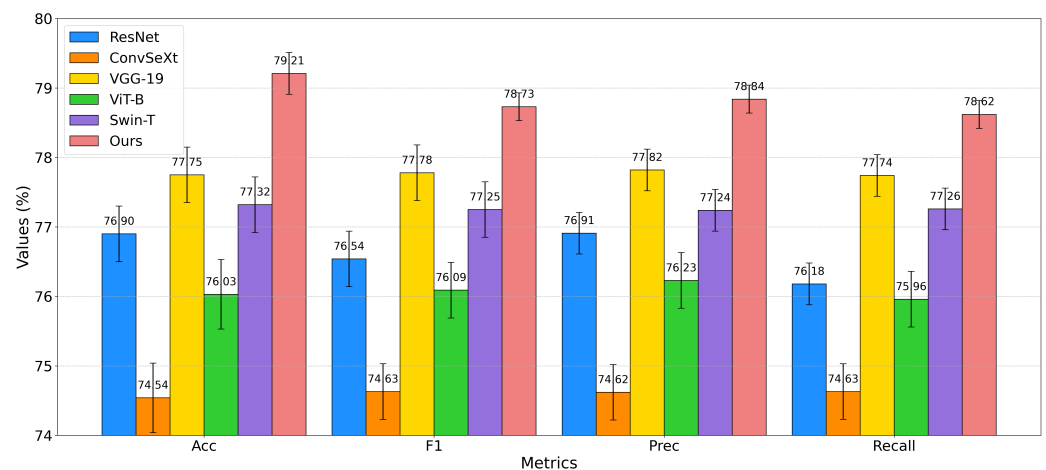
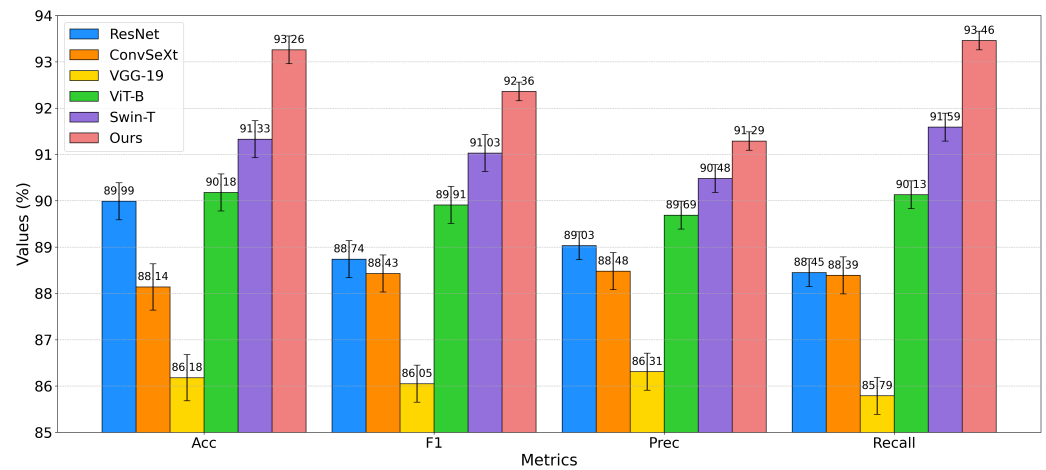
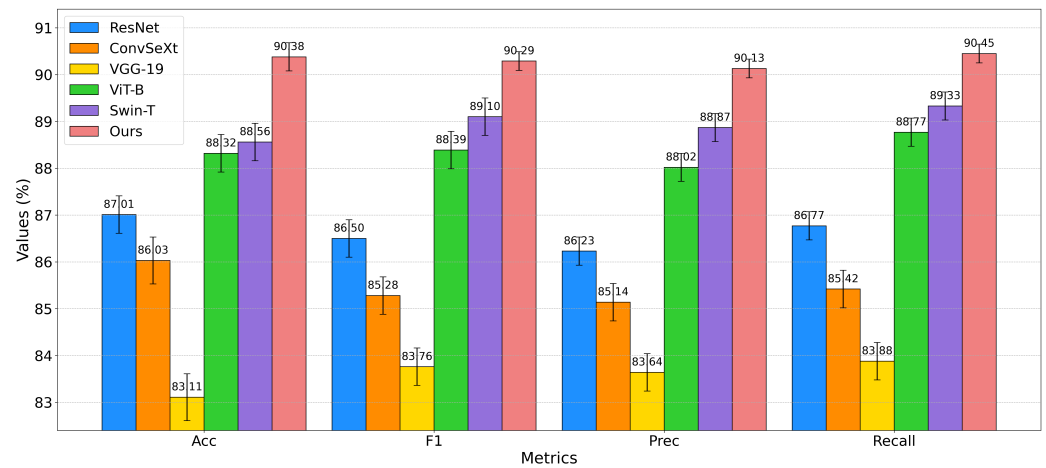


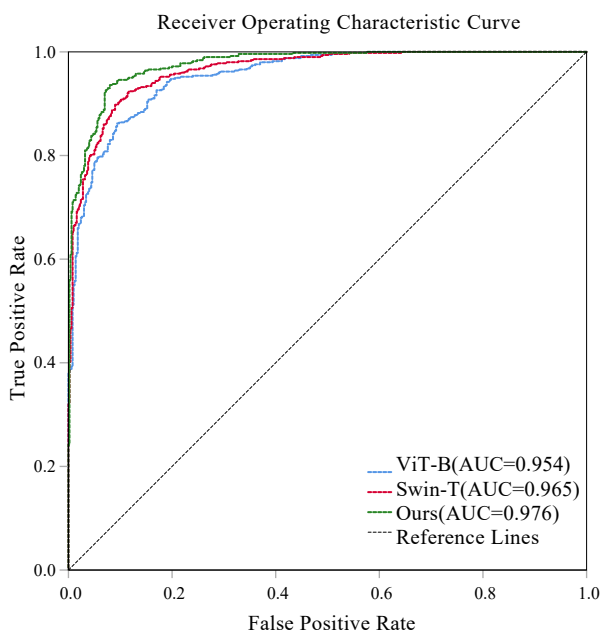
Figure 7. Visualization of performance between different algorithms on Kvasir dataset.



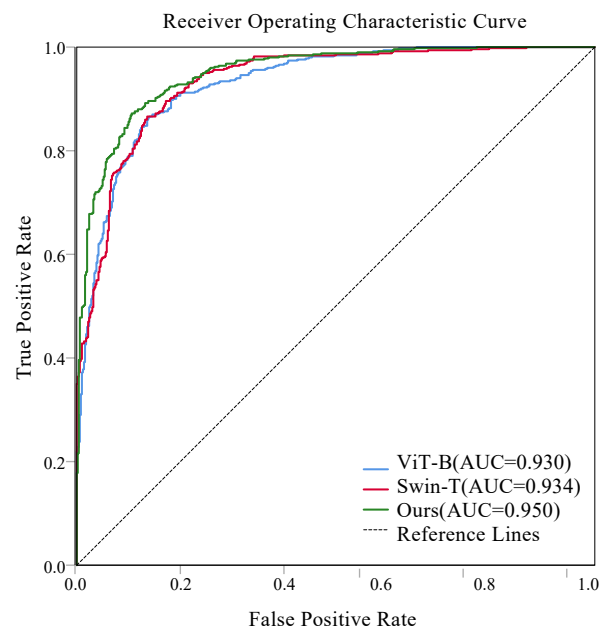
**Figure 8.** Visualization of performance between different algorithms on Brain Tumor Plus dataset.



**Figure 9.** Visualization of performance between different algorithms on Brain Tumor Plus-6 dataset.



**(a)** CPN-CX



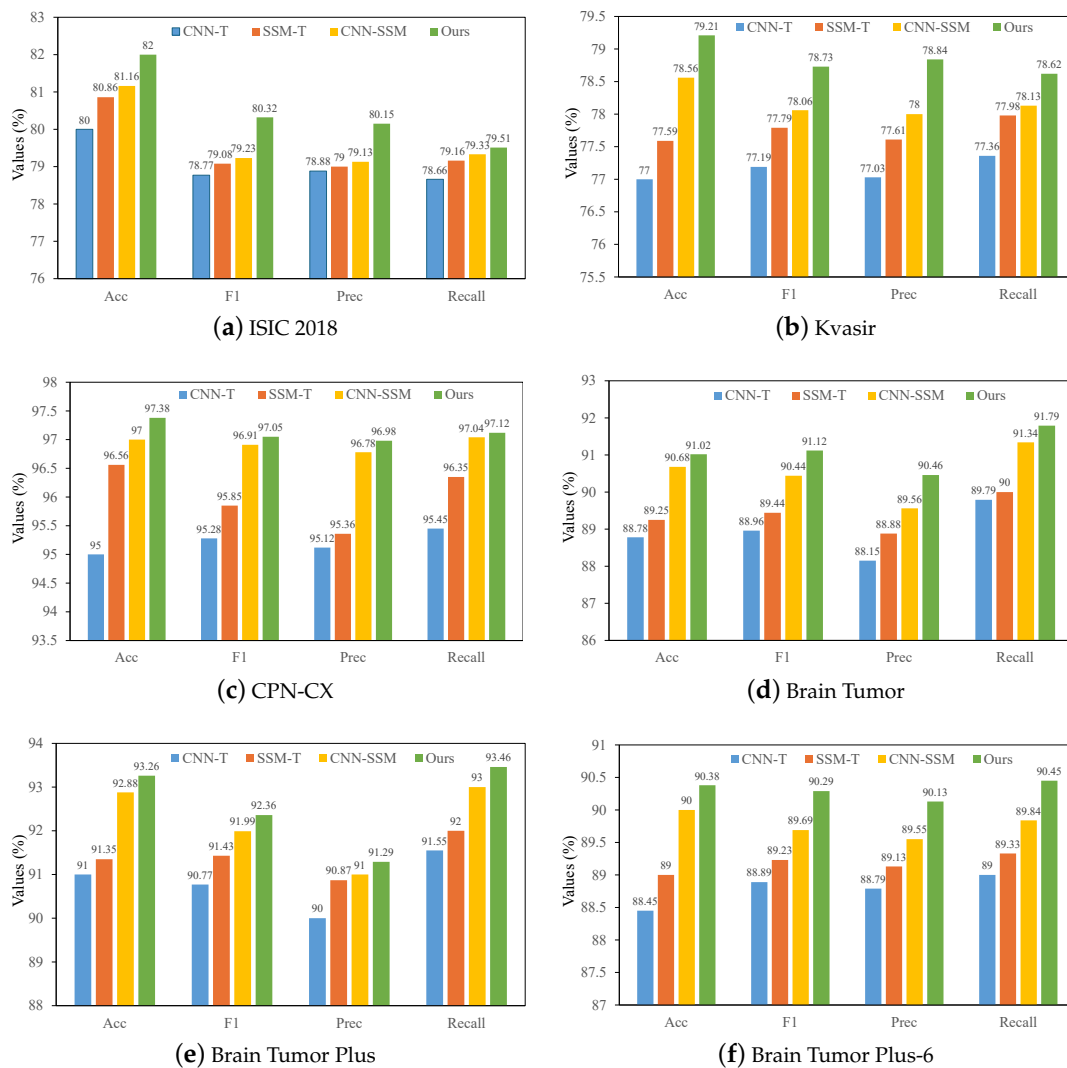
**(b)** Brain Tumor Plus

**Figure 10.** Average ROC curves for different methods.



#### 4.6. Ablation Studies

To further validate the reliability of our model, a systematic ablation study was conducted, as shown in Figure 11. Our new model combining CNN and Transformer is denoted as CNN-T (blue bars); the one combining SSM and Transformer is labeled as SSM-T (orange bars); the parallel combination of CNN and SSM is represented by CNN-SSM (yellow bars); and our model is referred to as Ours (green bars). Experiments were conducted on six datasets, and the results showed that our model outperformed in accuracy, F1 score, precision, and recall, followed by the CNN-SSM model, then the SSM-T model, with the CNN-T model performing the worst. This indicates that the SSM model is effective, as it captures distant fine-grained features by establishing long-range dependencies, thus enhancing model performance. Meanwhile, the combination of SSM and CNN's local modeling ability can play a greater advantage, this is because in medical images, the details of lesions are usually small and independent of normal tissues, accurately capturing the local feature information is important for cognition of the image, while Transformer usually loses the local information when focusing on global contextual information, which leads to a lesser cognition of the details than CNN. So, the combination of SSM and CNN is superior to the combination of SSM and Transformer.



**Figure 11.** Performance of different models on different datasets.

## 5. Conclusions

This paper presents a MambaConvT model based on SSM and investigates its feasibility for medical image classification across different categories. To evaluate the model's performance in medical image classification tasks, extensive experiments were conducted, comparing it with state-of-the-art classification algorithms. The experimental results demonstrate that the model presented in this paper achieves superior performance in classification tasks, surpassing the latest classification algorithms in terms of accuracy, precision, recall, F1 score, and AUC values. Moreover, the ablation studies confirm that the SSM module of our algorithm can enhance the model's performance by capturing distant detail features through the establishment of long-range dependencies. This approach offers a new perspective for CAD systems based on images, potentially expanding their applications. Despite the promising performance of the model in medical image classification tasks, this paper does not delve into the relationship between the inference speed and parameter size of the MambaConvT model. Additionally, the suitability of the model for high-resolution images and more advanced medical image tasks warrants further investigation.

**Author Contributions:** Conceptualization, H.Z. and C.H.; methodology, C.H.; software, C.H.; validation, C.H. and N.C.; formal analysis C.H., N.C. and B.G.; resources C.H. and N.C.; writing—original draft preparation, C.H.; writing—review and editing, C.H.; funding acquisition, N.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Jiangsu Provincial Key Research and Development Program BE2020714. Funded by Cao Ning for 100,000 RMB.

**Data Availability Statement:** We sincerely thank the International Skin Imaging Collaboration for releasing the ISIC open dataset available at <https://challenge.isic-archive.com/landing/2018/> accessed on 28 July 2024, the hospitals under the Vestre Viken Health Trust in Norway for providing the open dataset Kvasir available at <https://paperswithcode.com/dataset/kvasir> accessed on 28 July 2024, and the open datasets CPN-CX available at <https://paperswithcode.com/dataset/kvasir> accessed on 28 July 2024 and Brain Tumor available at <https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri> accessed on 28 July 2024 provided at the Kaggle Challenge available at <https://www.kaggle.com/> accessed on 28 July 2024. We also thank the Department of Radiology, Jiangsu Provincial People's Hospital, our collaborating hospital, for providing private datasets, which are available upon request from the corresponding author.

**Conflicts of Interest:** The authors declare no competing interests.

## References

1. Wang, W.; Liang, D.; Chen, Q.; Iwamoto, Y.; Han, X.H.; Zhang, Q.; Hu, H.; Lin, L.; Chen, Y.W. Medical image classification using deep learning. In *Deep Learning in Healthcare: Paradigms and Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 33–51.
2. Afshar, P.; Heidarian, S.; Enshaei, N.; Naderkhani, F.; Rafiee, M.J.; Oikonomou, A.; Fard, F.B.; Samimi, K.; Plataniotis, K.N.; Mohammadi, A. COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. *Sci. Data* **2021**, *8*, 121. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Serena Low, W.C.; Chuah, J.H.; Tee, C.A.T.; Anis, S.; Shoaib, M.A.; Faisal, A.; Khalil, A.; Lai, K.W. An Overview of Deep Learning Techniques on Chest X-Ray and CT Scan Identification of COVID-19. *Comput. Math. Methods Med.* **2021**, *2021*, 5528144. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Zhang, D.; Huang, G.; Zhang, Q.; Han, J.; Han, J.; Wang, Y.; Yu, Y. Exploring task structure for brain tumor segmentation from multi-modality MR images. *IEEE Trans. Image Process.* **2020**, *29*, 9032–9043. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Zhang, D.; Huang, G.; Zhang, Q.; Han, J.; Han, J.; Yu, Y. Cross-modality deep feature learning for brain tumor segmentation. *Pattern Recognit.* **2021**, *110*, 107562. [\[CrossRef\]](#)
6. Hu, R.; Li, M.; Xu, H.; Zhao, H.M. Research and application of key technologies for medical image intelligence knowledge discovery and data processing. *Int. J. Pattern Recognit. Artif. Intell.* **2020**, *34*, 2057005. [\[CrossRef\]](#)
7. Guan, X.; Ni, B.; Zhang, J.; Zhu, D.; Cai, Z.; Meng, W.; Shi, L.; Ross-Degnan, D. Association between physicians' workload and prescribing quality in one tertiary hospital in China. *J. Patient Saf.* **2021**, *17*, e1860–e1865. [\[CrossRef\]](#)
8. Tan, T.; Platel, B.; Huisman, H.; Sánchez, C.I.; Mus, R.; Karssemeijer, N. Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation. *IEEE Trans. Med. Imaging* **2012**, *31*, 1034–1042. [\[CrossRef\]](#)
9. Song, Y.; Cai, W.; Zhou, Y.; Feng, D.D. Feature-based image patch approximation for lung tissue classification. *IEEE Trans. Med. Imaging* **2013**, *32*, 797–808. [\[CrossRef\]](#) [\[PubMed\]](#)

10. Zhang, F.; Song, Y.; Cai, W.; Lee, M.Z.; Zhou, Y.; Huang, H.; Shan, S.; Fulham, M.J.; Feng, D.D. Lung nodule classification with multilevel patch-based context analysis. *IEEE Trans. Biomed. Eng.* **2013**, *61*, 1155–1166. [[CrossRef](#)] [[PubMed](#)]
11. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
12. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [[CrossRef](#)]
13. Nanni, L.; Lumini, A.; Brahnam, S. Local binary patterns variants as texture descriptors for medical image analysis. *Artif. Intell. Med.* **2010**, *49*, 117–125. [[CrossRef](#)] [[PubMed](#)]
14. Chen, X.; Wang, X.; Zhang, K.; Fung, K.M.; Thai, T.C.; Moore, K.; Mannel, R.S.; Liu, H.; Zheng, B.; Qiu, Y. Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* **2022**, *79*, 102444. [[CrossRef](#)] [[PubMed](#)]
15. Gu, Y.; Chi, J.; Liu, J.; Yang, L.; Zhang, B.; Yu, D.; Zhao, Y.; Lu, X. A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning. *Comput. Biol. Med.* **2021**, *137*, 104806. [[CrossRef](#)] [[PubMed](#)]
16. Aljuaid, H.; Alturki, N.; Alsubaie, N.; Cavallaro, L.; Liotta, A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Comput. Methods Programs Biomed.* **2022**, *223*, 106951. [[CrossRef](#)] [[PubMed](#)]
17. Khalil, M.; Khalil, A.; Ngom, A. A Comprehensive Study of Vision Transformers in Image Classification Tasks. *arXiv* **2023**, arXiv:2312.01232.
18. Barzekar, H.; Yu, Z. C-Net: A reliable convolutional neural network for biomedical image classification. *Expert Syst. Appl.* **2022**, *187*, 116003. [[CrossRef](#)]
19. Kumar, K.A.; Prasad, A.; Metan, J. A hybrid deep CNN-Cov-19-Res-Net Transfer learning archetype for an enhanced Brain tumor Detection and Classification scheme in medical image processing. *Biomed. Signal Process. Control.* **2022**, *76*, 103631.
20. Salehi, A.W.; Khan, S.; Gupta, G.; Alabdullah, B.I.; Almjally, A.; Alsolai, H.; Siddiqui, T.; Mellit, A. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability* **2023**, *15*, 5930. [[CrossRef](#)]
21. Manzari, O.N.; Ahmadabadi, H.; Kashiani, H.; Shokouhi, S.B.; Ayatollahi, A. MedViT: A robust vision transformer for generalized medical image classification. *Comput. Biol. Med.* **2023**, *157*, 106791. [[CrossRef](#)]
22. Wu, X.; Feng, Y.; Xu, H.; Lin, Z.; Chen, T.; Li, S.; Qiu, S.; Liu, Q.; Ma, Y.; Zhang, S. CTransCNN: Combining transformer and CNN in multilabel medical image classification. *Knowl.-Based Syst.* **2023**, *281*, 111030. [[CrossRef](#)]
23. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* **2021**, arXiv:2111.00396.
24. Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; Ré, C. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 572–585.
25. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
26. Ma, J.; Li, F.; Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv* **2024**, arXiv:2401.04722.
27. Yue, Y.; Li, Z. MedMamba: Vision Mamba for Medical Image Classification. *arXiv* **2024**, arXiv:2403.03849.
28. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. Vmamba: Visual state space model. *arXiv* **2024**, arXiv:2401.10166.
29. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
30. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
31. Koonce, B.; Koonce, B.E. *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Springer: Berlin/Heidelberg, Germany, 2021.
32. Zhu, Y.; Newsam, S. Densenet for dense flow. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 790–794.
33. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16133–16142.
34. Ashwath, V.; Sikha, O.; Benitez, R. TS-CNN: A three-tier self-interpretable CNN for multi-region medical image classification. *IEEE Access* **2023**, *11*, 78402–78418. [[CrossRef](#)]
35. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
36. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [[CrossRef](#)] [[PubMed](#)]
37. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
39. Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; Ré, C. Hippo: Recurrent memory with optimal polynomial projections. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1474–1487.

40. Nguyen, E.; Goel, K.; Gu, A.; Downs, G.; Shah, P.; Dao, T.; Baccus, S.; Ré, C. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2846–2861.
41. Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; Zhu, L. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv* **2024**, arXiv:2401.13560.
42. Ruan, J.; Xiang, S. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv* **2024**, arXiv:2402.02491.
43. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
44. Pogorelov, K.; Randel, K.R.; Griwodz, C.; Eskeland, S.L.; de Lange, T.; Johansen, D.; Spampinato, C.; Dang-Nguyen, D.T.; Lux, M.; Schmidt, P.T.; et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In Proceedings of the 8th ACM on Multimedia Systems Conference, Taipei, Taiwan, 20–23 June 2017; pp. 164–169.
45. Bhuvaji, S.; Kadam, A.; Bhumkar, P.; Dedge, S.; Kanchan, S. Brain Tumor Classification (MRI) Dataset. 2020. Available online: <https://www.kaggle.com/sartajbhuvaji/brain-tumor-classification-mri> (accessed on 28 July 2024).
46. Wen, L.; Li, X.; Li, X.; Gao, L. A new transfer learning based on VGG-19 network for fault diagnosis. In Proceedings of the 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), Porto, Portugal, 6–8 May 2019; pp. 205–209.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.