MedConvMamba: Enhancing Medical Image Classification by Integrating Convolutional Neural Networks with Mamba for Local Feature Extraction and Global Context Awareness

Hui Qu, Jiaxi Liu, Mingchen Zou, Xingyu Li, Canwei Dong, Xiaoyu Cui. *Member, IEEE*

***Abstract*— Medical image classification is a critical task in computer vision, with Convolutional Neural Networks (CNNs) and Transformers being widely used. However, CNNs face challenges in capturing global context and long-range dependencies, while Transformers encounter computational difficulties due to their quadratic complexity. Recently, State Space Models (SSMs), such as Mamba, have made significant progress in efficiently modeling long-range interactions with linear complexity. Inspired by these advancements, we introduce MedConvMamba for medical image classification. Our proposed SS-Conv-SSM module combines the local feature extraction capabilities of convolutional layers with the long-range dependency modeling of SSMs, enabling the model to better understand the overall structure of images while accurately identifying local lesions, making it particularly suitable for medical image classification. We validated MedConvMamba through experiments on two different modalities of medical image datasets, and the results demonstrate that MedConvMamba effectively detects lesions in various medical images. Additionally, we explored the parameter configurations and module settings of the CNN and SSM combined model through a series of experiments. This study provides valuable insights for developing more efficient AI algorithms and medical application systems based on CNN and SSM.**

***Index Terms*—Medical Images, Image Classification, Deep Learning, State Space Models, Convolutional Neural Networks**

I. Introduction

Modern medical research cannot do without the support of medical images, which provide precise visual representations of the structure and function of various tissues and organs in the human body, helping medical professionals and researchers to explore in detail the normal and abnormal conditions in patients, thus serving clinical and research purposes [1]. Nowadays, both cutting-edge research in laboratories and disease diagnosis by clinical doctors heavily rely on the rich information provided by medical image analysis to promote scientific inference and diagnosis [2], [3]. With the development of medical technology, various medical imaging methods have emerged. In clinical settings, these technologies are widely used. However, the accuracy of detection and diagnosis for cancer and many other diseases depends on the professional knowledge of individual clinical doctors, which leads to significant differences in the interpretation and interpretation of medical images. To address this clinical challenge, many computer-aided detection and diagnosis (CAD) solutions have been developed and tested, aimed at helping clinical doctors more effectively read medical images and make diagnostic decisions in a more accurate and objective manner [4].

Deep learning technology has shown significant potential in the field of image processing, especially in the field of computer-aided detection (CAD), where it has become a mainstream technology [5]. This technology is widely used in tasks such as medical image classification, segmentation, and object detection, among which image classification is particularly crucial as it directly relates to the patient's diagnostic process [6]. In medical image classification, models based on Convolutional Neural Networks (CNN) and Transformers exhibit excellent capabilities [7], [8]. These models not only efficiently extract features from images, but are also commonly used as backbones, encoders, or decoders in various model architectures [9]. However, each of these technologies has its limitations. Due to the limitations of local receptive domains, CNN models are difficult to capture remote information in images, which may lead to insufficient feature extraction. Although Transformer based models perform well in global information modeling, their self-attention mechanism relies on computationally intensive operations. When processing large-sized medical images, their computational burden significantly increases, which limits their application in practical medical image processing tasks. In addition, the secondary complexity requirement of this model may make it difficult to implement in resource constrained environments, especially in clinical

This work was supported in part by the China Key Research and Development Program (Grant No. 2023YFC2508200), the Natural Science Foundation of Liaoning Province (Grant No.2022-MS-105) and the Liaoning Province Medical Engineering Cross Joint Fund (Grant No. 2022-YGJC-06).

Hui Qu, Jiaxi Liu, Mingchen Zou, Xingyu Li and Canwei Dong are with the College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110004, China
e-mail: quhui950328@163.com; 2271335@stu.neu.edu.cn; 2171286 @stu.neu.edu.cn; whiteli1925@outlook.com; 15978697657@163.com
Xiaoyu Cui is the corresponding author. He is with the College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110004, China and also with the Key Laboratory of Biomedical Imaging Science and SystemChinese Academy of Sciences (e-mail: cuixy@bmie.neu.edu.cn).

scenarios that require fast and efficient processing of large amounts of image data [7]–[10]. Therefore, although these technologies have been applied in multiple healthcare fields, they still need to be optimized for specific application scenarios to improve performance.

Based on these facts and the relatively complex nature of medical images compared to natural images, there is an urgent need to develop a new medical image classification architecture that can effectively capture remote information while maintaining linear computational complexity while achieving local feature extraction. In recent years, State Space Modeling (SSM) [11], [12] has aroused great interest among researchers. On the basis of classical SSM research, Modern SSMs such as Mamba not only establish long-range dependencies, but also exhibit linear complexity related to input size [13]. In addition, models based on SSM have been widely studied in many fields, including language comprehension [12], [13], general vision [14], [15], medical segmentation [16], [17], etc. Especially some studies use SSM to complete various image processing tasks, such as natural image classification and medical image segmentation. However, the performance of medical image classification models based on SSM still needs to be explored.

Inspired by the success of SSM in natural image classification tasks, we propose MedConvMamba, a lightweight model based on CNN and SSM, aimed at showcasing the potential of combining CNN and Mamba in various medical image classification tasks. Specifically, The core of MedConvMamba consists of a module called SConvs-SSM. Unlike natural images, various medical images have high similarity in overall image structure and strong heterogeneity in local lesions [18]. Therefore, the model needs to effectively extract local fine-grained features and global features. Based on this fundamental fact, SConvs-SSM combines the ability of convolutional layers to understand local context and the remote image context integration ability of SSM, while ensuring the lightweight of the model, achieving effective extraction of local and global features in medical images and accurate image classification.

The main contributions of this article are as follows: 1) We propose MedConvMamba and explore the potential applications of CNN combined with SSM models in different modalities of medical images. 2) A comprehensive experiment was conducted on two different modalities of datasets, and the results showed that MedConvMamba has considerable competitiveness while ensuring lightweight. 3) This study has established a new baseline for medical image classification tasks, providing valuable insights for the development of more efficient and effective artificial intelligence algorithms and medical application systems based on CNN combined with SSM in the future.

II. Methods

A. Preliminaries

The SSM-based models, such as structured state space sequence models (S4) and Mamba, are inspired by the continuous system [13]–[19], which maps a 1-D function or sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ through a $h(t) \in \mathbb{R}^N$ hidden state. This system uses $A \in \mathbb{R}^{N\times N}$ as the evolution parameter and $B \in \mathbb{R}^{N\times 1}, C \in \mathbb{R}^{1\times N}$ as the projection parameters.

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t). \end{aligned} \tag{1}$$

The S4 and Mamba are the discrete versions of the continuous system, which include a timescale parameter Δ to transform the continuous parameters A, B to discrete parameters $\bar{A}$, $\bar{B}$. The commonly used method for transformation is zero-order hold (ZOH), which is defined as follows:

$$\begin{aligned} \bar{A} &= \exp(\Delta A), \\ \bar{B} &= (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B. \end{aligned} \tag{2}$$

After the discretization of $\bar{A}$, $\bar{B}$, the discretized version of Eq. (1) using a step size Δ can be rewritten as:

$$\begin{aligned} h_t &= \bar{A}h_{t-1} + \bar{B}x_t, \\ y_t &= Ch_t. \end{aligned} \tag{3}$$

At last, the models compute output through a global convolution.

$$\begin{aligned} \bar{K} &= (C\bar{B}, C\bar{A}B, \ldots, CA^{M-1}\bar{B}), \\ y &= x * \bar{K}, \end{aligned} \tag{4}$$

where M is the length of the input sequence x, and $\bar{K} \in \mathbb{R}^M$ is a structured convolutional kernel.

B. MedConvMamba

Figure 1 shows the overall architecture of MedConvMamba. The structure of MedConvMamba is quite simple, comprising a patch embedding layer, SConvs-SSM Block, and an embedding fusion module.

In MedConvMamba, the patch embedding layer first divides the input image $x \in \mathbb{R}^{H\times W\times 3}$ into non-overlapping patches of size 8×8, mapping the image dimensions to C (The default value of C is 192 for the Tiny version of the model and 384 for the Small version) , resulting in the embedded image $x' \in \mathbb{R}^{(H/8)\times(W/8)\times C}$. Before feeding it into the backbone of MedConvMamba, x' is normalized using layer normalization [20]. The backbone consists of several SConvs-SSM Blocks. The output dimension of each module is the same as x, that is, $y_i \in \mathbb{R}^{(H/8)\times(W/8)\times C}$. y_i is the output of the i-th SConvs-SSM Block. The output of the last SConvs-SSM block will be input into the embedding fusion module to obtain the final prediction.

1) Mamba Architecture: Mamba is a simplified SSM architecture that differs from the stacking method of traditional linear attention modules and multi-layer perceptron (MLP) [21] modules. It integrates these two modules into a single Mamba module. This module simplifies the structure by replacing multiplication gating with activation functions and integrating SSM transformation into the main path of MLP. The overall design of Mamba includes multiple such modules, equipped with standard normalization layers and residual connections at intervals. This architecture not only

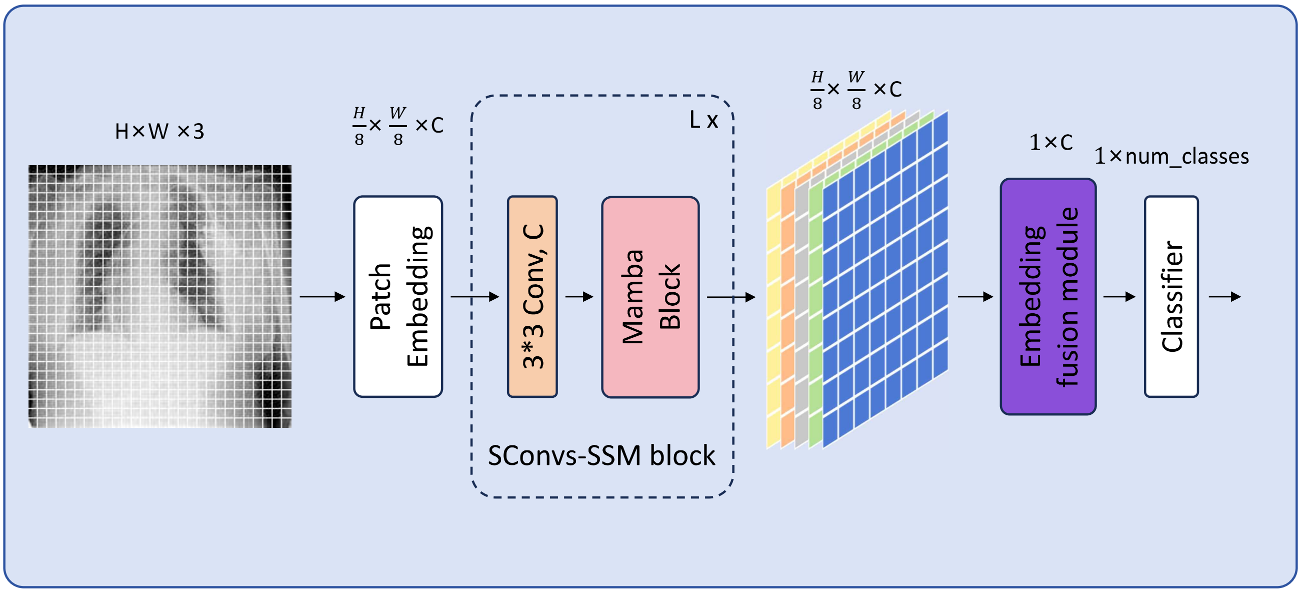

Fig. 1. The overall architecture of the MedConvMamba.

maintains the linear extension ability of the state space model to sequence length, but also has the modeling effect of a transformer. Mamba combines the advantages of CNN and transformers, providing an efficient foundational model for the field of computational vision (CV). Unlike traditional converters, Mamba avoids explicitly storing the entire context through a selection mechanism, and the one-dimensional and causal characteristics of this mechanism are its research focus in CV applications.

2) SConvs-SSM block: The SConvs-SSM block is the core module of MedConvMamba, as shown in Figure 2. This block is designed to be a simple yet effective module, combining a single convolutional layer with a Mamba structure. The pseudo-code for the Mamba structure is presented in Algorithm 1. The simplicity of the SConvs-SSM block not only streamlines the architecture but also reduces the number of parameters, making the model more lightweight and efficient. Firstly, the input is processed through a convolutional layer with a kernel size of 3, a step size of 1, and a C-channel. Then input the output of this layer into the Mamba architecture. In Mamba, we linearly project the normalized sequence to the x and z with dimension size E. Then, x is first subjected to one-dimensional convolution. Subsequently, the embedding is linearly projected onto B, C and Δ. Δ is used to transform A and B. The output result is calculated by SSM, selected by z, and combined to produce an output that maintains the same shape as the module input [22], [23]. This approach not only optimizes the representation of features but also enhances the model's performance in handling classification tasks. The SConvs-SSM block's main advantage is its ability to combine the robust local feature extraction of CNNs with the Mamba structure's expanded receptive field. This integration enhances medical imaging by capturing fine-grained anomalies and contextual information, improving classification accuracy. As shown in Figure 3, the SConvs SSM block retains the good local feature extraction ability of traditional CNN while achieving a larger receptive field than Transformer, effectively balancing local detail focus and a wider range of image backgrounds.

Algorithm 1 Pseudo-code for Mamba in SConvs-SSM block

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

3) Embedding fusion module: The embedding fusion module receives the output from the last SConvs-SSM block. Its primary function is to integrate the same embedding features across all patches and compute a unified value for each type of embedding, thereby generating a $1 \times C$ feature vector. This process is achieved through global average pooling, effectively consolidating information from various features. This feature vector is then fed into a fully connected layer, which is responsible for the final classification prediction.

III. Experiments and Results

In this section, we conducted experiments with MedConvMamba, applying it to two types of medical image classification tasks: X-ray images and dermatoscopic images (Figure 4).

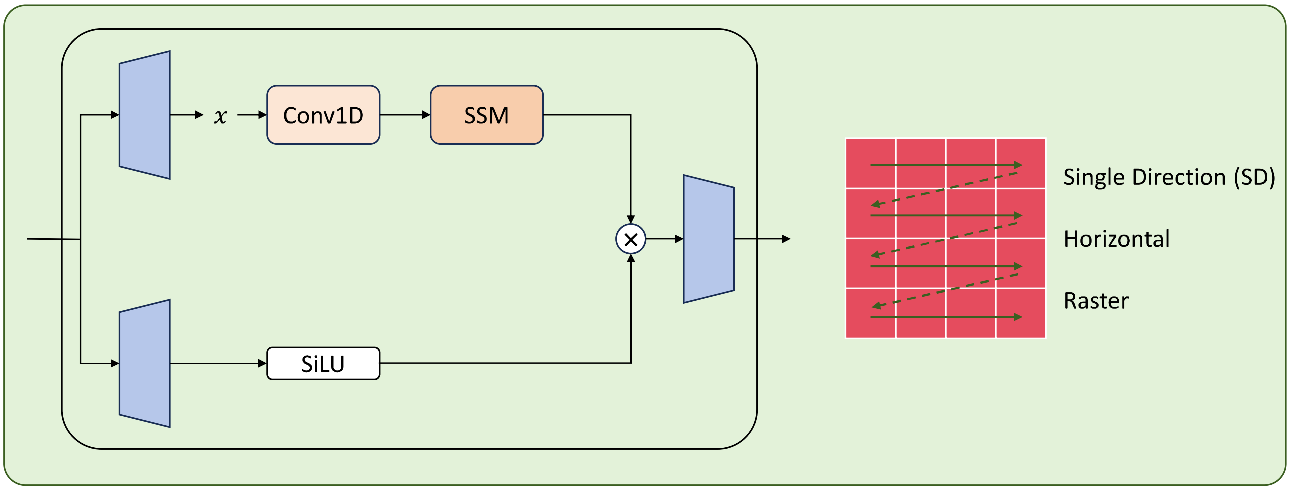

Fig. 2. Architecture of Mamba block. Mamba combines the base blocks of SSM with the MLP blocks prevalent in modern neural networks to form a new Mamba block, which is stacked and combined with normalization and residual connection to form the Mamba network architecture.

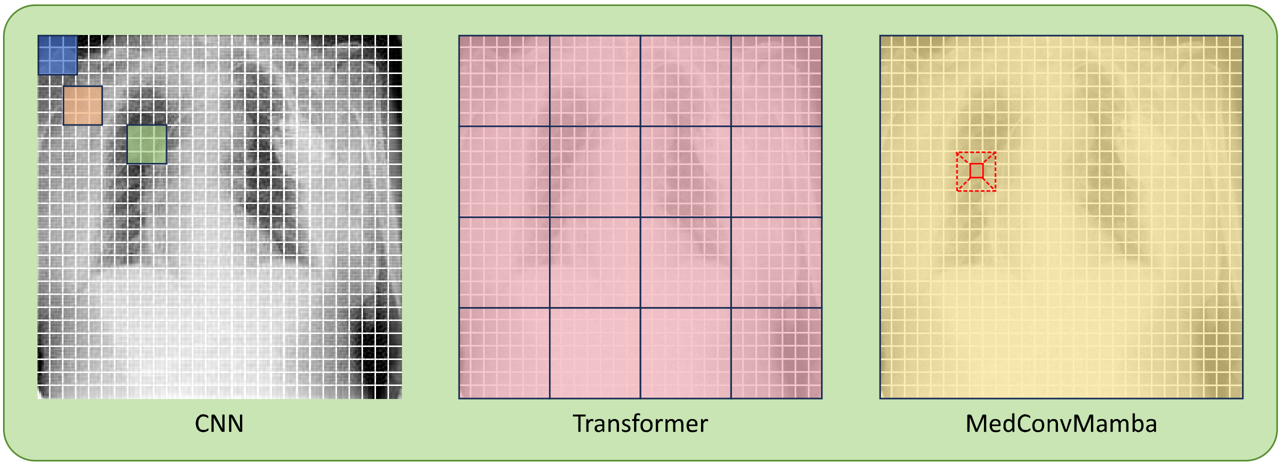

Fig. 3. Respective field comparison among CNN, Transformer, and our proposed MedConvMamba.

A. Embedding fusion module

COVID-19 Radiography Database The COVID-19 Radiography Database [24] is a publicly available medical imaging database, primarily jointly published by the University of Qatar, Dalhousie University, and ARK Information Solutions in India. This database focuses on collecting chest X-ray images related to COVID-19 to help researchers and doctors better understand and diagnose the COVID-19 virus through image analysis. The latest version includes 3616 COVID-19 positive cases, as well as 10192 normal, 6012 pulmonary opacities (non COVID pulmonary infections), and 1345 images of viral pneumonia.

The ISIC2019 [25] dataset comprises 25,331 dermoscopic images, collected from various international sources, including the Medical University of Vienna (MUV), Austria, and several clinics in Queensland, Australia. These images have been gathered over the past three decades. In the earlier years, before digital cameras were common, lesion images were captured on film, stored, and archived at MUV's Department of Dermatology. These analog photographs were later digitized using a high-resolution scanner, converted into 8-bit color JPEG format at a resolution of 300 DPI, and subsequently resized to a resolution of 72 DPI with dimensions of 1024 × 768 pixels. The dataset is categorized into eight different classes, namely vascular lesions (VASC), actinic keratosis (AKIEC), melanoma (MEL), benign keratosis (BKL), melanocytic nevus (NV), basal cell carcinoma (BCC), dermatofibroma (DF), and squamous cell carcinoma (SCC), containing 193, 435, 1323, 1242, 12100, 678, 239, and 221 images respectively.

B. Evaluation Metrics

This study used five key indicators – accuracy (ACC), precision, recall, specificity, and F1 score - to evaluate the classification performance of the model. These indicators are widely recognized and used in medical image classification research by calculating the true examples (TP), true negative examples (TN), false negative examples (FN), and false positive examples (FP) in the confusion matrix. The above

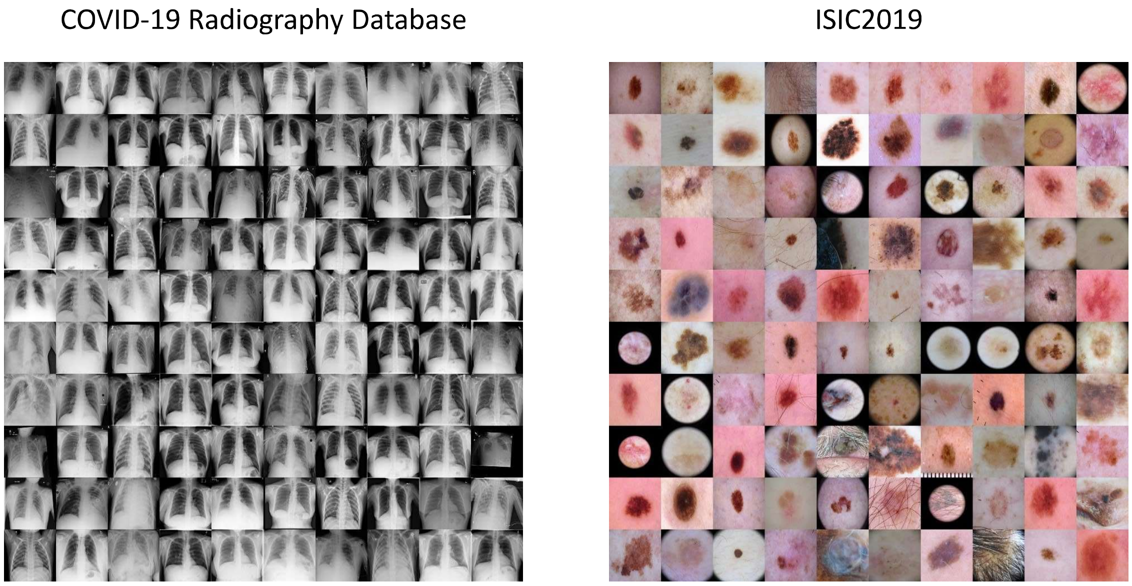

Fig. 4. Respective field comparison among CNN, Transformer, and our proposed MedConvMamba.

evaluation indicators are calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{8}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{9}$$

C. Implementation Details

Before initiating the network training, all images were resized to dimensions of 224×224×3. Each image underwent normalization and standardization processes. We used the AdamW optimizer [26] for adjusting model parameters, setting an initial learning rate of 5e-4, with a weight decay of 1e-5, and employed Cross-Entropy Loss. The model was trained using the PyTorch [27] framework over 300 epochs with a batch size of 64. To prevent overfitting, an early-stop strategy was implemented. No data augmentation techniques or pretrained weights were used, ensuring that the performance metrics of the model could directly reflect the unique architecture of MedConvMamba. The training setup included an Ubuntu 18.04 operating system and an NVIDIA RTX A6000 GPU.

D. Results

1) Comparing Classification Performance with Other Models: In previous work, researchers have tended to use classification models meticulously designed for ImageNet [28] to classify medical images. Although these models were initially applied to natural image classification, experimental results have shown that they are also effective in medical image classification [29], [30]. Therefore, to demonstrate the potential of MedConvMamba, we compared it with state-of-the-art Mamba-based medical image classification models, as well as with some of the most advanced models applied to medical image classification tasks. The experimental results are shown in Tables I and II.

The results indicate that, despite the simplicity of the MedConvMamba architecture, it performs excellently in medical image classification tasks. MedConvMamba-S surpasses mainstream Mamba-based models, as well as CNN and Transformer-based models, in terms of ACC on the two selected datasets. Notably, its sensitivity and F1 scores also exceed those of other models. In medical image classification tasks, sensitivity and F1 scores are crucial metrics. This is because, in multi-class disease diagnosis models, there is usually a greater emphasis on capturing diseased samples. Misdiagnosis in medical diagnostics (i.e., falsely classifying diseased samples as healthy) can lead to serious consequences, thus models need to minimize such errors. The superior performance of MedConvMamba in these two metrics further demonstrates its potential in medical image classification tasks.

2) The Impact of the Number of Blocks on Model Performance: In the SConvs SSM block, we utilized a simplified Mamba architecture as the SSM component in the module. Vim increases model depth by using 24 Vim blocks similar to the Mamba architecture. Although VMamba uses VSS blocks with more parameters, it uses fewer blocks.

TABLE I
COMPARISON OF CLASSIFICATION PERFORMANCE OF DIFFERENT MODELS ON COVID-19 RADIOGRAPH DATABASE

	Image size	Param	GFLOPs	Acc	Precision	Sensitivity	Specificity	F1 Score
CNN								
ResNet18 [31]	224	11M	1.8	81.11	82.45	75.35	92.23	78.23
ResNet34 [31]	224	22.7M	3.7	81.25	82.59	76.67	92.33	78.98
RegNetY-4G [32]	224	19.6M	4.0	93.36	93.91	93.85	**97.30**	93.86
ConvNext [33]	224	28.6M	4.5	89.75	92.52	87.90	95.58	89.98
Vit								
Vit-T [34]	224	4.9M	1.3	87.30	87.10	86.99	94.76	86.97
Swin-T [35]	224	28.3M	4.5	92.26	92.37	92.26	94.57	92.23
DeiT-Ti [36]	224	5.7M	1.3	88.81	90.75	87.10	95.35	88.76
Efficientvit_M0 [37]	224	2.3M	0.08	92.42	94.02	90.53	96.53	92.14
Efficientvit_M4 [37]	224	8.8M	0.3	93.55	**94.83**	92.04	96.89	93.33
SSM								
VMamba-T [14]	224	30.7M	4.9	94.15	94.10	94.03	97.47	94.02
MedMamba [38]	224	13.3M	2.0	93.77	93.88	93.77	97.56	93.80
Vim-T [15]	224	7.1M	1.5	91.90	93.66	90.66	96.57	92.07
Vim-S [15]	224	26.0M	5.1	92.96	94.47	92.34	96.99	93.35
MedConvMamba-T	224	4.1M	1.8	93.60	93.62	93.60	95.98	93.60
MedConvMamba-S	224	16.1M	7.3	**94.45**	94.47	**94.45**	96.35	**94.45**

TABLE II
COMPARISON OF CLASSIFICATION PERFORMANCE OF DIFFERENT MODELS ON ISIC2019 DATABASE

	Image size	Param	GFLOPs	Acc	Precision	Sensitivity	Specificity	F1 Score
CNN								
ResNet18	224	11M	1.8	65.02	37.25	30.97	93.09	31.49
ResNet34	224	22.7M	3.7	63.87	30.45	29.69	92.99	29.78
RegNetY-4G	224	19.6M	4.0	70.47	56.95	53.42	94.77	54.68
ConvNext	224	28.6M	4.5	68.95	62.49	46.62	93.88	49.84
Vit								
Vit-T	224	4.9M	1.3	67.28	57.89	60.52	92.59	57.84
Swin-T	224	28.3M	4.5	77.75	76.96	77.75	89.57	76.91
DeiT-Ti	224	5.7M	1.3	71.83	65.45	49.09	94.69	52.63
Efficientvit_M0	224	2.3M	0.08	66.86	49.07	29.67	92.78	30.47
Efficientvit_M4	224	8.8M	0.3	69.21	47.93	34.94	93.43	37.30
SSM								
VMamba-T	224	30.7M	4.9	75.92	74.32	75.50	95.42	74.31
MedMamba	224	13.3M	2.0	75.20	74.70	75.18	95.52	74.37
Vim-T	224	7.1M	1.5	72.20	68.34	53.04	94.73	57.35
Vim-S	224	26.0M	5.1	76.65	72.84	62.39	**95.66**	66.55
MedConvMamba-T	224	4.1M	1.8	76.39	75.26	76.39	88.44	75.36
MedConvMamba-S	224	16.1M	7.3	**79.69**	**79.33**	**79.69**	90.13	**79.03**

To determine the optimal depth setting for MedConvMamba, we conducted experiments with varying depths based on MedConvMamba-T using the COVID-19 Radiograph Dataset. As shown in Table III, maintaining other parameters constant, appropriately reducing the depth not only decreased the number of model parameters but also improved performance. However, excessively shallow depths significantly reduced accuracy. Therefore, based on the experimental results, the depth of the model in MedConvMamba was set to 7.

TABLE III
PERFORMANCE OF MEDCONVMAMBA-T AT DIFFERENT DEPTHS ON COVID-19 RADIOGRAPH DATASET

Depth	Image size	Param	GFLOPs	Acc
3	224	1.8M	0.8	92.8
7	224	**4.1M**	1.8	**93.6**
14	224	8.2M	3.7	86.5

3) The Impact of Patch Size Settings on Model Performance: In previous studies, Vim followed the ViT architecture and set the patch size to 16. However, in medical imaging, lesions can be quite small, necessitating the model's ability to detect fine-grained details. Therefore, a smaller patch size is required to capture more detailed features. Additionally, a smaller patch size increases the input sequence length. The Mamba architecture's advantage with long sequence inputs makes it more suitable for medical image classification compared to Transformers. As shown in Table IV, reducing the patch size improved the predictive performance of MedConvMamba-T on COVID-19 Radiograph dataset. Consequently, in MedConvMamba-T, the patch size was set to 8.

TABLE IV
PERFORMANCE OF MEDCONVMAMBA-T ON RADIOGRAPH DATASETS OF COVID-19 UNDER DIFFERENT PATCH SIZE SETTINGS

Patch size	Image size	Param	ACC	Acc
8	224	4.1M	93.6	92.8
14	224	4.2M	92.6	93.6
16	224	4.2M	92.0	86.5

4) Comparison of Different Embedding Fusion Methods: For embedding fusion, we experimented with four different methods and the results are shown in Table V. Initially, we employed the simplest approach by flattening all features into a one-dimensional vector and then passing it through a fully connected layer to generate the feature vector. This method achieved excellent classification performance with a relatively small number of parameters. Next, we used a standard convolutional kernel with a size of 3x3 and a stride of 1 to convolve the feature maps, followed by global average pooling to generate the feature vector. This approach significantly increased the number of parameters and resulted in a slight decrease in performance. Subsequently, we replaced the standard convolution with depth-wise convolution [39], which yielded the best performance while maintaining a lower parameter count compared to the previous methods. Finally, when we removed the depth-wise convolution layer and only used global average pooling to generate the feature vector, the parameter count was further reduced without degrading the model's performance. Therefore, in MedConvMamba, we employ global average pooling as the embedding fusion module to extract the feature vector.

TABLE V
PERFORMANCE OF MEDCONVMAMBA-T ON RADIOGRAPH DATASETS OF COVID-19 UNDER DIFFERENT PATCH SIZE SETTINGS

	Image size	Param	GFLOPs	Acc
Flatten	224	4.7M	1.9	93.5
Convolution	224	33M	1.9	93.2
Depth-wise convolution	224	4.3M	1.8	93.6
Global average pooling	**224**	**4.1M**	1.8	**93.6**

IV. CONCLUSION

In this study, to effectively model medical images with fine-grained features, we combined CNNs, which are adept at extracting local features, with SSMs, which can handle long sequence inputs with fewer parameters to capture global features. This led to the development of MedConvMamba, a model tailored for medical image classification tasks. To demonstrate that the potential of MedConvMamba in medical classification tasks directly benefits from SSM, we did not employ any pretraining strategies or data augmentation techniques. We conducted experiments on MedConvMamba using two different types of medical datasets and performed several experiments to optimize the parameter settings for the application of SSM in the medical imaging domain. The results show that MedConvMamba exhibits excellent performance in medical image classification tasks. Moreover, compared to various widely used architectures in medical classification tasks, MedConvMamba is highly competitive.

For future work, we outline the following points: 1) We will further explore the potential of LightMed Mamba on medical datasets obtained from other imaging technologies and optimize it, with a focus on lightweight design and suitability for smaller datasets. 2) We plan to introduce causal inference [40] concepts and methods to enhance the interpretability of MedConvMamba, making it more competitive in the medical field. Additionally, developing visualization methods suitable for SSM is necessary for the same purpose. 3) Given the advantages of SSM in long-sequence modeling, it is worthwhile to investigate the application prospects of MedConvMamba in high-resolution medical images, such as pathology images [41]. 4) Using MedConvMamba as a backbone, we will explore its potential applications in other medical imaging tasks, such as image segmentation, object detection, image registration, and image reconstruction.

REFERENCES

[1] S. Suganyadevi, V. Seethalakshmi, K. Balasamy, A review on deep learning in medical image analysis, International Journal of Multimedia Information Retrieval 11(1) (2022) 19-38.
[2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature medicine 25(1) (2019) 24-29.
[3] J. Wang, H. Zhu, S.-H. Wang, Y.-D. Zhang, A review of deep learning on medical image analysis, Mobile Networks and Applications 26(1) (2021) 351-380.
[4] H. Ayesha, S. Iqbal, M. Tariq, M. Abrar, M. Sanaullah, I. Abbas, A. Rehman, M.F.K. Niazi, S. Hussain, Automatic medical image interpretation: State of the art and future directions, Pattern Recognition 114 (2021) 107856.
[5] Y.J. Cha, R. Ali, J. Lewis, O. Büyüköztürk, Deep learning-based structural health monitoring, Automation in Construction 161 (2024).
[6] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T.C. Thai, K. Moore, R.S. Mannel, H. Liu, B. Zheng, Y. Qiu, Recent advances and clinical applications of deep learning in medical image analysis, Medical Image Analysis 79 (2022) 102444.
[7] F. Shamshad, S. Khan, S.W. Zamir, M.H. Khan, M. Hayat, F.S. Khan, H. Fu, Transformers in medical imaging: A survey, Medical Image Analysis (2023) 102802.
[8] S.S. Kshatri, D. Singh, Convolutional neural network in medical image analysis: A review, Archives of Computational Methods in Engineering 30(4) (2023) 2793-2810.
[9] O. Elharrouss, Y. Akbari, N. Almadeed, S. Al-Maadeed, Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision, Computer Science Review 53 (2024) 100645.
[10] G. Papanastasiou, N. Dikaios, J. Huang, C. Wang, G. Yang, Is attention all you need in medical image analysis? A review, IEEE Journal of Biomedical and Health Informatics (2023).
[11] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, C. Ré, Combining recurrent, convolutional, and continuous-time models with linear state space layers, Advances in neural information processing systems 34 (2021) 572-585.

[12] A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, arXiv preprint arXiv:2111.00396 (2021).

[13] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023).

[14] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: Visual state space model, arXiv preprint arXiv:2401.10166 (2024).

[15] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, arXiv preprint arXiv:2401.09417 (2024).

[16] Z. Xing, T. Ye, Y. Yang, G. Liu, L. Zhu, Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation, arXiv preprint arXiv:2401.13560 (2024).

[17] J. Ma, F. Li, B. Wang, U-mamba: Enhancing long-range dependency for biomedical image segmentation, arXiv preprint arXiv:2401.04722 (2024).

[18] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, M. De Bruijne, Semi-supervised medical image segmentation via learning consistency under transformations, Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, Springer, 2019, pp. 810-818.

[19] J. Ruan, S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, arXiv preprint arXiv:2402.02491 (2024).

[20] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[21] M.-C. Popescu, V.E. Balas, L. Perescu-Popescu, N. Mastorakis, Multilayer perceptron and neural networks, WSEAS Transactions on Circuits and Systems 8(7) (2009) 579-588.

[22] X. Liu, C. Zhang, L. Zhang, Vision Mamba: A Comprehensive Survey and Taxonomy, arXiv preprint arXiv:2405.04404 (2024).

[23] R. Xu, S. Yang, Y. Wang, B. Du, H. Chen, A survey on vision mamba: Models, applications and challenges, arXiv preprint arXiv:2404.18861 (2024).

[24] M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K. Reajul Islam, M. Salman Khan, A. Iqbal, N. Al-Emadi, Can AI help in screening Viral and COVID-19 pneumonia?, arXiv (2020).

[25] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), arXiv preprint arXiv:1902.03368 (2019).

[26] I. Loshchilov, F. Hutter, Fixing Weight Decay Regularization in Adam, (2017).

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248-255.

[29] Y. Wen, L. Chen, Y. Deng, C. Zhou, Rethinking pre-training on medical imaging, Journal of Visual Communication and Image Representation 78 (2021) 103145.

[30] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Computers in biology and medicine 128 (2021) 104115.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[32] I. Radosavovic, R.P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10428-10436.

[33] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976-11986.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012-10022.

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers distillation through attention, International conference on machine learning, PMLR, 2021, pp. 10347-10357.

[37] H. Cai, J. Li, M. Hu, C. Gan, S. Han, Efficientvit: Lightweight multi-scale attention for on-device semantic segmentation, arXiv preprint arXiv:2205.14756 (2022).

[38] Y. Yue, Z. Li, Medmamba: Vision mamba for medical image classification, arXiv preprint arXiv:2403.03849 (2024).

[39] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[40] Y. Luo, J. Peng, J. Ma, When causal inference meets deep learning, Nature Machine Intelligence 2(8) (2020) 426-427.

[41] S. Banerji, S. Mitra, Deep learning in histopathology: A review, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12(1) (2022) e1439.