GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Medicine and Life Sciences

Article

Research on Medical Image Segmentation Algorithm Based on a Lightweight Attention Convolutional Neural Network

Xinzhe Yuan [1,*]

[1] Northeastern University, San Jose, California, United States
* Correspondence: Xinzhe Yuan, Northeastern University, San Jose, California, United States

Abstract: With the continuous development of medical imaging technology, medical image segmentation is playing an increasingly important role in clinical diagnosis and treatment planning. However, traditional deep learning methods, while ensuring segmentation accuracy, often suffer from issues such as large model size and high computational complexity. To address these challenges, this paper proposes a medical image segmentation algorithm based on a lightweight attention convolutional neural network. By incorporating lightweight convolution modules (such as depthwise separable convolutions and group convolutions), the proposed algorithm effectively reduces the number of model parameters and computational burden. At the same time, it integrates attention mechanisms — including channel attention and spatial attention — to enhance feature representation, thereby achieving higher accuracy and robustness across various medical image segmentation tasks. Experiments conducted on several public datasets, in comparison with mainstream methods, demonstrate significant advantages in both segmentation precision and operational efficiency. The research presented in this paper provides new ideas and references for the development of lightweight medical image segmentation techniques.

Keywords: medical image segmentation; lightweight network; attention mechanism; Convolutional Neural Network; deep learning

Received: 28 February 2025
Revised: 10 March 2025
Accepted: 25 March 2025
Published: 30 March 2025

1. Introduction

Medical image segmentation is essential for computer-aided diagnosis and treatment planning. With advances in imaging devices and an explosion of data, accurately extracting target organs or lesions from complex backgrounds has become critical for improving clinical outcomes. While traditional segmentation methods have achieved notable accuracy and robustness, they often face challenges such as large model sizes, high computational complexity, and insufficient real-time performance — issues that are especially problematic in resource-constrained settings. To address these challenges, we propose a segmentation algorithm based on a lightweight attention convolutional neural network. Built on an encoder–decoder framework, our approach incorporates lightweight designs like depthwise separable and group convolutions to significantly reduce the network's parameter count and computational burden. Additionally, attention mechanisms are employed to weight features, enhancing the network's ability to capture critical lesion regions and anatomical details. This combination achieves high-precision segmentation across multiple modalities while maintaining fast inference speeds. Experimental results on several datasets show high IoU and Dice coefficients, confirming the method's effectiveness for clinical applications. Overall, our work provides a balanced solution for efficient and accurate medical image segmentation, paving the way for the broader deployment of intelligent medical technologies in clinical practice and suggesting new directions for future research [1].

2. Medical Image Segmentation Techniques: Concepts and Applications

2.1. Overview of Medical Image Segmentation Techniques

Medical image segmentation is crucial for clinical diagnosis and treatment planning, aiming to accurately extract regions of interest — such as organs or lesions — from complex backgrounds. This process provides precise information for pathological analysis, surgical planning, and treatment evaluation. With advances in imaging equipment and techniques, modalities like X-ray, CT, MRI, and ultrasound now offer rich data with high resolution, but also present significant challenges in algorithm design. Different imaging modalities have unique characteristics. For instance, X-ray images primarily show the outlines of bones and high-density tissues, CT scans provide detailed tissue density information, MRI excels at soft tissue contrast, and ultrasound images are often affected by noise and artifacts. Consequently, segmentation algorithms must be both general and adaptable enough to leverage the unique features of each modality [2].

Figure 1. outlines a typical workflow for medical image segmentation. First, the original image is acquired based on clinical needs (e.g., musculoskeletal or chest X-rays). Next, the region of interest (ROI) is localized by focusing on key areas such as bone structures or lung fields. Then, pixel-level segmentation is performed to generate binary or multi-class masks that highlight the anatomical structures or lesions. Finally, ground truth images — annotated manually or semi-automatically — are used for training, validation, and evaluation. In musculoskeletal X-rays, segmentation focuses on bones or joints for fracture detection and arthritis assessment, while chest X-rays require precise delineation of lung fields and lesions for early disease detection [3]. The segmentation process not only minimizes background interference but also accentuates important features, facilitating further quantitative analysis and 3D reconstruction. Deep learning, particularly convolutional neural networks (CNNs), has made significant strides in this field by providing end-to-end learning and robust feature extraction. However, traditional large-scale CNNs face challenges such as high computational demands, excessive parameters, and dependence on hardware resources — issues that hinder real-time clinical use. This has led to a growing interest in lightweight network architectures using depthwise separable and group convolutions to reduce computational and storage overhead without compromising accuracy. Simultaneously, attention mechanisms have shown promise by dynamically weighting channels or spatial regions, thereby enhancing the focus on key features even in noisy or complex images. Combining lightweight designs with attention mechanisms is a current research focus aimed at maximizing segmentation performance while minimizing resource consumption. Overall, medical image segmentation has evolved from traditional methods to high-precision deep learning models. Ongoing innovations continue to address the challenges posed by diverse medical images and clinical demands, supporting the advancement of precision medicine and intelligent healthcare [4].

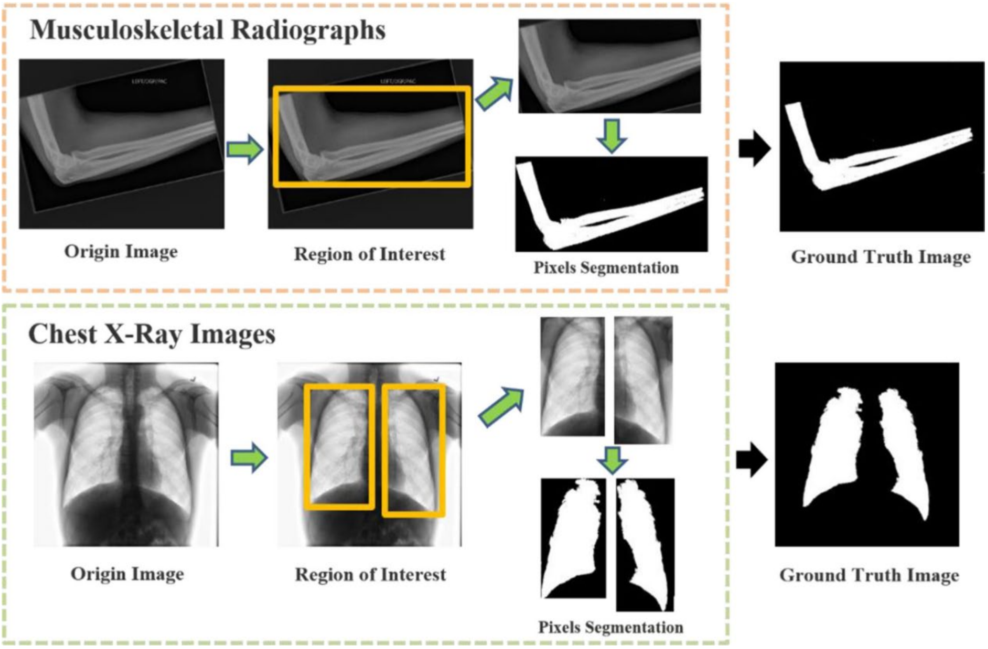

Figure 1: Example Workflow for Segmenting Different Types of X-ray Images

2.2. Application of the Lightweight Attention Convolutional Neural Network in Medical Image Segmentation

In medical image segmentation, a lightweight attention convolutional neural network effectively balances high accuracy with low computational and storage overhead. As shown in Figure 2, the model is built on an encoder–decoder framework that integrates lightweight convolution and attention modules. This integration allows the network to efficiently capture key features and suppress redundant information across various medical imaging scenarios. The network begins by passing the input image through a series of convolution blocks and pooling operations, which progressively reduce spatial resolution while extracting multi-level semantic features. Unlike traditional convolutional networks, the lightweight design uses strategies such as depthwise separable convolutions and group convolutions to significantly lower the number of parameters and computational load [5]. Batch normalization and activation functions are applied to ensure stable training and prevent redundancy from excessive operators. Following these convolutional layers, the feature maps are processed by several attention blocks. These blocks employ global average pooling and fully connected layers to re-weight feature channels or spatial positions, emphasizing critical features while diminishing irrelevant or noisy regions. This results in robust segmentation performance, even with complex backgrounds and different imaging modalities. The decoder then restores spatial resolution through upsampling and feature fusion, using skip connections to retain fine details from earlier layers for richer segmentation outputs [6].

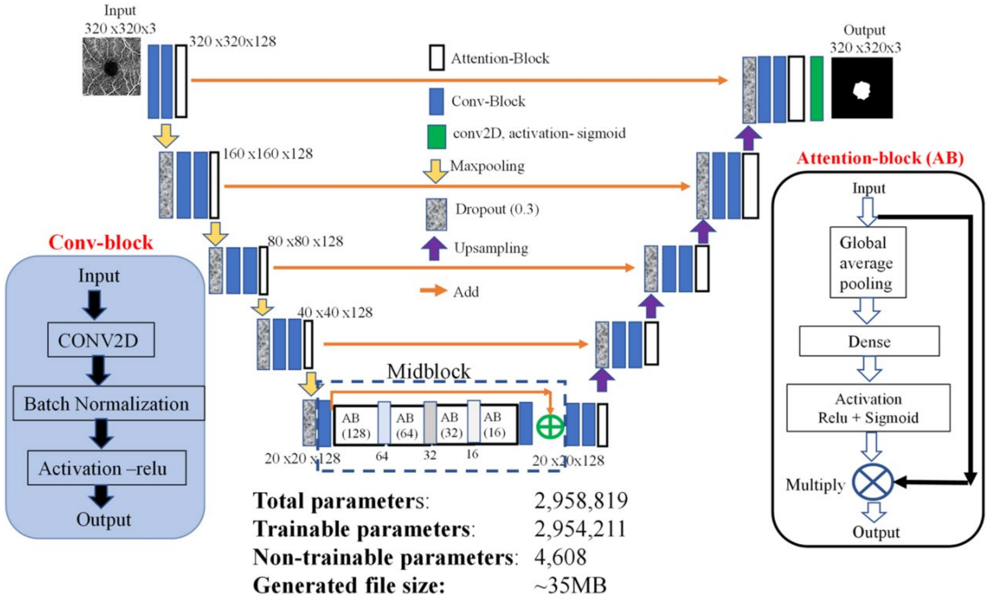

Figure 2. Illustration of the overall architecture of a lightweight attention convolutional neural network.

Figure 2 illustrates that the full network comprises about 2.95 million trainable parameters, with non-trainable parameters around 4608 and a model file size of approximately 35 MB. Compared with conventional large-scale segmentation networks, this lightweight attention network is ideally suited for deployment in resource-constrained environments, such as portable medical devices or real-time applications. Its design achieves a balance between efficiency and accuracy, making it highly adaptable to various segmentation tasks including identification of bones, lung fields, hearts, and other organs or lesions, thus providing strong technical support for intelligent healthcare [7].

3. Methodology

3.1. Overall Framework Design

In medical image segmentation, attention mechanisms are crucial for enabling networks to focus on key features while keeping computational demands low. Unlike traditional recurrent or convolution operations, attention mechanisms allow flexible allocation of computational resources by adaptively weighting important regions and suppressing noise. This enhances feature representation and improves the precise localization of target tissues or lesions within a lightweight structure [8].

Figure 3 illustrates these benefits. Part (a) shows RNNsearch-style attention that progressively focuses on key content by weighting sequence context, whereas part (b) depicts self-attention in image processing. Self-attention establishes mappings among queries (Q), keys (K), and values (V) to capture global features directly, without relying on sequential scanning. This facilitates parallel computation and better handles long-distance dependencies, which is particularly beneficial for identifying complex lesions and widespread feature correlations. In our proposed lightweight attention convolutional neural network, the attention mechanism is applied between the encoder and decoder to re-weight feature maps. By incorporating self-attention modules, the network adaptively selects and amplifies important features along both the channel and spatial dimensions while extracting high-level semantic information. This approach maintains high segmentation accuracy even in complex imaging scenarios and supports fast inference due to its non-recurrent structure. Moreover, the lightweight design and attention mechanisms complement each other [9]. Lightweight convolutions — such as depthwise separable and group convolutions — significantly reduce the number of parameters and computational load, making it feasible to run deep segmentation models on resource-constrained hardware. Meanwhile, self-attention modules boost the network's ability to capture critical information with minimal parameter increase. Together, these strategies effectively balance segmentation accuracy and efficiency, unlocking the network's potential for clinical diagnosis and treatment planning. Subsequent sections will detail performance evaluations on various datasets and present ablation experiments that validate the contributions of both lightweight convolution and attention mechanisms [10].

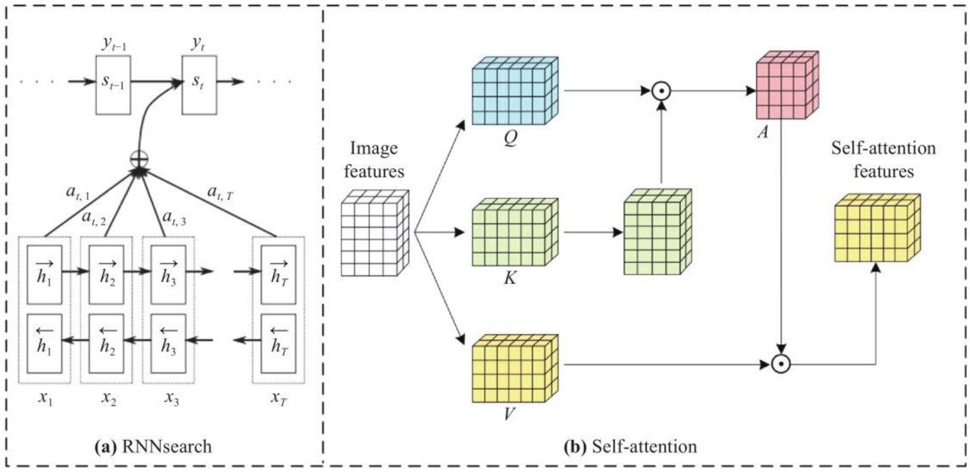

Figure 3. Schematic Diagram of Feature Extraction Comparison Based on Sequence Modeling and Self-Attention.

3.2. Lightweight Module Design

Large deep convolutional networks offer excellent feature extraction but often have high parameter counts and computational costs that hinder real-time clinical applications. To overcome this, our study integrates various lightweight convolution strategies to reduce computational overhead while maintaining or improving segmentation accuracy. We primarily employ depthwise separable convolutions and group convolutions. Depthwise separable convolutions split a standard convolution into a depthwise and a pointwise convolution, significantly reducing the number of parameters and multiply-add operations. Group convolutions partition input features into groups and perform separate convolutions, lowering inter-channel redundancy for more efficient feature extraction. Both methods decrease dependency on computational resources, enabling high efficiency even on mobile devices or low-compute environments. In our network, both the encoder and decoder are constructed using lightweight convolutional units. Compared to traditional convolution layers, these units dramatically reduce trainable parameters and storage overhead, and are paired with batch normalization and nonlinear activations (e.g., ReLU or Leaky ReLU) to ensure stable training and balanced feature distributions. In the encoder, lightweight convolutions focus on quickly compressing spatial dimensions and extracting basic features, while the decoder balances the upsampling of high-level semantic information with restoring low-level details. By using depthwise separable or group convolutions on both ends, our approach reduces overall parameters while preserving image detail capture, thereby supporting effective attention-based feature re-weighting later. Additionally, skip connections and residual structures are integrated at key positions to facilitate gradient propagation and mitigate training challenges from increased depth. Regularization techniques like dropout further suppress overfitting, enhancing generalization. Overall, the lightweight module design is not just about reducing load but achieving a balanced trade-off between efficiency and segmentation accuracy, making our network a feasible and flexible solution for clinical medical image segmentation.

3.3. Loss Functions and Optimization Strategy

In medical image segmentation tasks, selecting an appropriate loss function is critical for enhancing both the training effectiveness and segmentation accuracy of the model. Given that medical images often involve imbalanced data (e.g., lesion regions may occupy a much smaller area compared to normal tissues) and require high precision, a single loss function is often insufficient to capture both global and local features adequately. Therefore, this study incorporates multiple forms of loss functions — such as cross-entropy and Dice coefficient — and further reinforces focus on the target regions by applying weighting or combined strategies. Concurrently, in terms of optimization strategy, adaptive gradient descent methods (such as Adam) or stochastic gradient descent (SGD) with momentum are employed to achieve stable convergence within a relatively short training period. For clarity, the following are three typical loss function formulas: weighted cross-entropy loss, Dice loss, and a combined loss that fuses both. In practical applications, the weighting coefficients can be adjusted or additional regularization terms added depending on the data distribution and segmentation requirements. Firstly, to address the issue of class imbalance, this study applies a weighting scheme to the traditional cross-entropy loss. For each class cc, a weight wcw_c is assigned, and the cross-entropy loss is defined as shown in Formula 1:

$$L_{CE} = -\sum_{c=1}^{C} w_c \, y_c \log(\hat{y}_c) \tag{1}$$

Where C is the number of classes, y_c denotes the one-hot encoded ground truth label, $\hat{y}_c$ is the network's predicted output, and w_c balances the influence of each class in the loss calculation. When a certain class is underrepresented in the training set or is particularly critical for segmentation quality, its weight can be increased appropriately. Secondly, to emphasize the accurate capture of target region boundaries and mitigate class imbalance, a Dice coefficient-related loss is often employed. Let pip_i represent the prediction for the i-th pixel, gig_i the corresponding ground truth label, N the total number of pixels, and ϵ a smoothing term. The Dice loss is defined as shown in Formula 2:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon} \tag{2}$$

This loss function is very common in medical image segmentation as it directly reflects the degree of overlap between the predicted and ground truth regions. Finally, to combine the advantages of weighted cross-entropy and Dice loss, this study fuses the two through a weighted sum, resulting in a composite loss function as shown in Formula 3:

$$L = \alpha L_{CE} + \beta L_{Dice} \tag{3}$$

Where α and β are hyperparameters used to balance the contributions of the weighted cross-entropy and Dice losses in the overall optimization objective. By tuning these coefficients on different stages or datasets, the model can achieve an optimal trade-off between edge contour recognition and overall accuracy. Regarding the optimization strategy, this study primarily adopts adaptive optimization algorithms based on gradient descent (such as Adam) combined with momentum and learning rate decay strategies to accelerate convergence and reduce oscillations. Alternatively, stochastic gradient descent (SGD) with momentum or RMSProp can be used, depending on the data scale and network architecture. During training, monitoring the loss on the validation set and evaluation metrics (such as Dice coefficient and IoU) allows for early stopping or adjustment of the learning rate to obtain a more stable model performance and prevent overfitting. Through the careful design and combination of these loss functions and optimization strategies, the model is able to achieve higher segmentation accuracy and robustness when facing various types of medical images.

4. Experimental Design and Results Analysis

4.1. Dataset Introduction and Preprocessing

In this study, we utilized multiple publicly available medical image datasets to thoroughly validate the performance of the proposed lightweight attention convolutional neural network for segmentation tasks. Specifically, three representative medical imaging datasets were selected, covering chest X-ray images, abdominal CT images, and skin lesion images. Each dataset includes rich annotation information, which is suitable for training deep learning models as well as for quantitative evaluation of segmentation performance. Table 1 summarizes the key characteristics of each dataset, including the number of images, resolution, image type, annotation details, and data source.

Table 1. Overview of Datasets.

Dataset Name	Number of Images	Image Resolution	Image Type	Annotation Details	Data Source
Chest X-ray Dataset	1200	1024 × 1024	X-ray	Annotations for lung fields, heart, and lesion areas	Public clinical database
Abdominal CT Dataset	800	512 × 512	CT	Annotations for liver, spleen, kidneys, and other organs	Open medical imaging resource platform
Skin Lesion Dataset	1000	768 × 768	Digital image	Segmentation annotations for lesion areas	International skin lesion competition dataset

For data preprocessing, a standardized procedure was applied to each dataset to ensure consistency and improve the model's generalization capability. First, images were normalized, mapping pixel values to the [0,1] range. Second, based on the original resolutions, images were resized or cropped to a fixed size to facilitate batch training. For the abdominal CT and skin lesion images, data augmentation techniques such as random rotation, translation, scaling, and horizontal flipping were also employed to increase sample diversity and reduce overfitting. Additionally, to address the noise issues commonly found in medical images, certain datasets underwent filtering and contrast enhancement operations, which further improved the algorithm’s ability to capture details and edge information. Through these preprocessing steps, high-quality, balanced, and representative training and validation datasets were constructed, providing a solid foundation for the subsequent experiments.

4.2. Experimental Settings

4.2.1. Experimental Platform and Training Parameters

The experimental training for this study was conducted on a high-performance computing server with GPU acceleration to improve computational efficiency. All models were run under the same hardware and software environment to ensure comparability of results. Table 2 details the configuration of the experimental platform.

Table 2. Experimental Platform Configuration.

Component	Specification
Processor (CPU)	Intel Xeon Gold 6226R @ 2.90GHz (16 cores)
Graphics Card (GPU)	NVIDIA RTX A6000 (48GB)
Memory (RAM)	256GB DDR4
Operating System	Ubuntu 20.04 LTS

Deep Learning Framework	TensorFlow 2.8 / PyTorch 1.12
CUDA Version	11.3
cuDNN Version	8.2
Python Version	3.8

During training, hyperparameters were optimized to ensure that the model converged within a reasonable timeframe and achieved optimal segmentation performance. The specific training parameters are listed in Table 3.

Table 3. Training Hyperparameters.

Parameter	**Value**
Input Image Size	320 × 320 × 3
Batch Size	16
Initial Learning Rate	0.001
Learning Rate Decay	Cosine Annealing
Number of Epochs	100
Optimizer	Adam
Loss Function	Weighted Cross-Entropy + Dice Loss
Weight Initialization	Xavier Initialization
Regularization	L2 Regularization + Dropout (0.3)

A cosine annealing strategy was employed to dynamically adjust the learning rate, providing a high learning rate in the early stages to speed up convergence and reducing it later to prevent oscillations and overfitting. Additionally, Xavier initialization was used to ensure numerical stability during the initial training phase.

4.2.2. Evaluation Metrics

To objectively assess the segmentation performance of the proposed model, several common evaluation metrics were employed, including Intersection over Union (IoU), Dice Similarity Coefficient (DSC), Accuracy (ACC), and Sensitivity (SEN). IoU measures the overlap between the predicted segmentation and the ground truth, while the Dice coefficient emphasizes the consistency between the prediction and the annotation. Accuracy evaluates the overall classification correctness, and sensitivity reflects the model's ability to recall target regions.The evaluation metrics are defined by the following formulas 4-7:

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{4}$$

$$\mathrm{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{5}$$

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

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

Here, P and G represent the predicted segmentation mask and the ground truth mask, respectively; TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. During the experiments, the model was tested on the three different medical image datasets, and the primary evaluation metrics were recorded as shown in Figure 4.

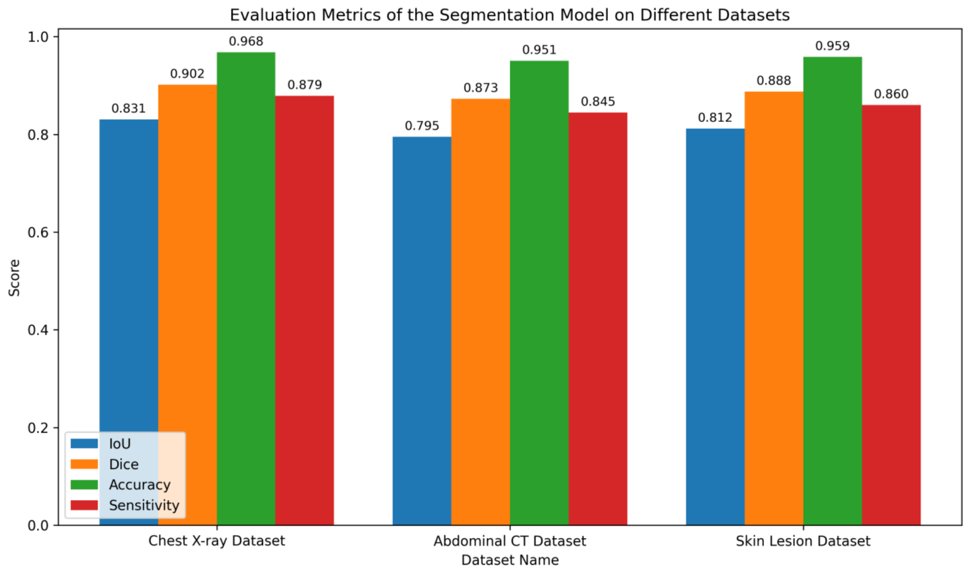

Figure 4. Evaluation Metrics of the Segmentation Model on Different Datasets.

The results indicate that the proposed lightweight attention convolutional neural network achieves high IoU and Dice coefficients across all datasets, demonstrating superior segmentation accuracy in extracting target regions. Moreover, the accuracy and sensitivity metrics suggest that the model generalizes well across different types of lesion segmentation tasks. Overall, the proposed medical image segmentation algorithm provides an efficient, accurate, and scalable solution for medical image analysis while significantly reducing computational complexity.

4.3. Experimental Results

In this section, we comprehensively evaluate the performance of the proposed lightweight attention convolutional neural network on medical image segmentation tasks from multiple perspectives, including segmentation accuracy, the contribution of different modules, and computational efficiency, to validate the effectiveness and practicality of our method. First, in terms of segmentation accuracy, the model was tested on the chest X-ray, abdominal CT, and skin lesion datasets, and the IoU, Dice, Accuracy, and Sensitivity metrics were recorded. As shown in Figure 5, the model achieved high segmentation accuracy across all three datasets. Specifically, the Chest X-ray Dataset achieved an IoU of 0.831 and a Dice coefficient of 0.902; the Abdominal CT Dataset achieved an IoU of 0.795 and a Dice coefficient of 0.873; and the Skin Lesion Dataset achieved values of 0.812 and 0.888, respectively. These results clearly indicate that the model consistently extracts the target regions from various medical images, exhibiting high overlap and overall accuracy. Next, to verify the specific contributions of the lightweight modules and attention mechanisms to the model's performance, ablation experiments were conducted. In these experiments, the attention module was removed, lightweight convolutions were replaced with standard convolutions, and both components were removed simultaneously, with key segmentation metrics recorded for each configuration. As shown in Figure 5, removing the attention module alone resulted in an average decrease in the Dice coefficient of approximately 2.5%. Replacing lightweight convolutions with standard convolutions significantly increased the parameter count, with only a slight improvement in segmentation accuracy; when both components were removed, the model's performance deteriorated noticeably. These results indicate that, while maintaining a lightweight model is important, the attention mechanism plays an indispensable role in enhancing segmentation accuracy.

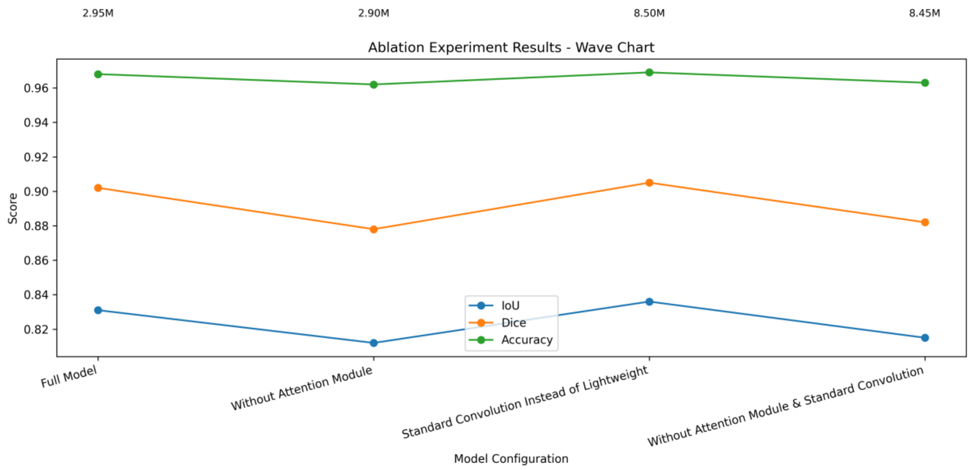

Figure 5. Ablation Experiment Results.

Finally, to evaluate the computational efficiency of the model for real-world deployment, the resource consumption of different model configurations was assessed, including inference time and model size. As demonstrated in Figure 6, the full model maintained high accuracy while having a significantly lower parameter count and smaller file size compared to the version built with standard convolutions, and the inference time was also reduced. This confirms that the lightweight design strategy adopted in this study can substantially lower computational resource requirements in practical applications, meeting the demands of real-time segmentation tasks.

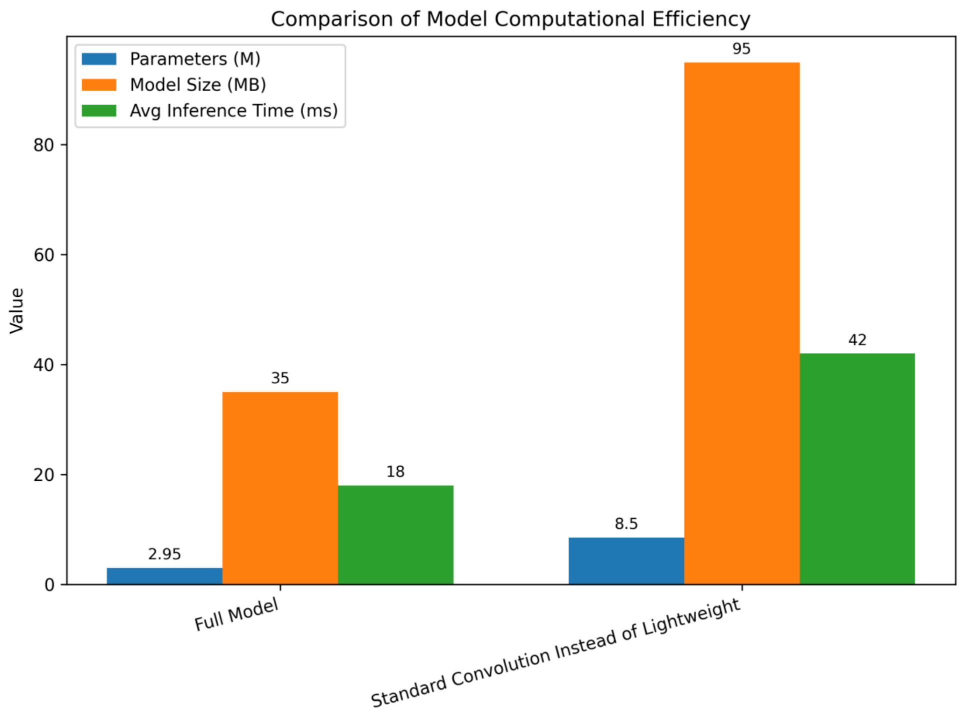

Figure 6. Comparison of Model Computational Efficiency.

In summary, the experimental results from multiple angles — segmentation accuracy, ablation studies, and computational efficiency — demonstrate that the proposed lightweight attention convolutional neural network exhibits excellent performance in medical image segmentation tasks. The model achieves high IoU and Dice coefficients across different datasets, and by integrating lightweight design with attention mechanisms, it successfully reduces parameter count and computational complexity while maintaining or even enhancing segmentation performance. These advantages make the model highly promising for practical clinical applications.

5. Conclusion

This paper presents a medical image segmentation algorithm based on a lightweight attention convolutional neural network. By integrating lightweight convolution modules with attention mechanisms, the proposed method effectively reduces the number of model parameters and computational complexity while achieving outstanding segmentation performance on multiple medical image datasets. Experimental results indicate that the method maintains high IoU and Dice coefficients and significantly improves inference speed, making it suitable for real-time clinical applications in resource-constrained environments. Future work will focus on further optimizing the lightweight design, enhancing the model's adaptability to complex lesion structures, and exploring multi-task learning strategies for cross-modal medical image segmentation to advance the deployment of intelligent medical technologies.

References

1. X. Li, Y. Jiang, M. Li, and S. Yin, "Lightweight attention convolutional neural network for retinal vessel image segmentation," *IEEE Trans. Ind. Inform.*, vol. 17, no. 3, pp. 1958–1967, Mar. 2020, doi; 10.1109/TII.2020.2993842.
2. J. Chen, W. Chen, A. Zeb, and D. Zhang, "Segmentation of medical images using an attention embedded lightweight network," *Eng. Appl. Artif. Intell.*, vol. 116, p. 105416, Feb. 2022, doi: 10.1016/j.engappai.2022.105416.
3. V. K. Singh, et al., "Prior wavelet knowledge for multi-modal medical image segmentation using a lightweight neural network with attention guided features," *Expert Syst. Appl.*, vol. 209, p. 118166, May 2022, doi: 10.1016/j.eswa.2022.118166.
4. Q. Zhou, Q. Wang, Y. Bao, L. Kong, X. Jin, and W. Ou, "LAEDNet: A lightweight attention encoder–decoder network for ultrasound medical image segmentation," *Comput. Electr. Eng.*, vol. 99, p. 107777, Jul. 2022, doi: 10.1016/j.compeleceng.2022.107777.
5. Q. Zhou, Z. Huang, M. Ding, and X. Zhang, "Medical image classification using lightweight CNN with spiking cortical model-based attention module," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 1991–2002, Apr. 2023, doi: 10.1109/JBHI.2023.3241439.
6. H. Liu, G. Huo, Q. Li, X. Guan and M. L. Tseng, "Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation," *Expert Syst. Appl.*, vol. 214, p. 119166, Sep. 2023, doi: 10.1016/j.eswa.2022.119166.
7. S. Iqbal, T. M. Khan, S. S. Naqvi, A. Naveed, M. Usman, H. A. Khan, and I. Razzak., "LDMRes-Net: A lightweight neural network for efficient medical image segmentation on IoT and edge devices," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 7, pp. 3860–3871, Jul. 2023, doi: 10.1109/JBHI.2023.3331278.
8. Z. Han, M. Jian, and G.-G. Wang, "ConvUNeXt: An efficient convolution neural network for medical image segmentation," *Knowl.-Based Syst.*, vol. 253, p. 109512, Oct. 2022, doi: 10.1016/j.knosys.2022.109512.
9. T. Agrawal and P. Choudhary, "ALCNN: Attention-based lightweight convolutional neural network for pneumothorax detection in chest X-rays," *Biomed. Signal Process. Control*, vol. 79, p. 104126, Jan. 2023, doi:10.1016/J.BSPC.2022.104126.
10. X. Lin, L. Yu, K. T. Cheng, and Z. Yan, "BATFormer: Towards boundary-aware lightweight transformer for efficient medical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 7, pp. 3501–3512, Jul. 2023, doi: 10.1109/JBHI.2023.3266977.