Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 1, January 2026, pp: 222-239 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

HALF-MAFUNET: A Lightweight Architecture Based on Multi-Scale Adaptive Fusion for Medical Image Segmentation

Abiaz Fazel Maula Sandy, Heri Prasetyo

Department of Informatics, Universitas Sebelas Maret, Surakarta, Indonesia

Corresponding author: Heri Prasetyo. (e-mail: heri.prasetyo@staff.uns.ac.id), **Author(s) Email**: Abiaz Fazel Maula Sandy (abiazfazel_ms@student.uns.ac.id)

Medical image segmentation is a critical component in computer-aided diagnosis systems but many deep learning models still require large numbers of parameters and heavy computation. Classical CNN-based architectures such as U-Net and its variants achieve good accuracy, but are often too heavy for real deployment. Meanwhile, modern Transformer-based or Mamba-based models capture long-range information but typically increase model complexity. Because of these limitations, there is still a need for a lightweight segmentation model that can provide a good balance between accuracy and efficiency across different types of medical images. This paper proposes Half-MAFUNet, a lightweight architecture based on multi-scale adaptive fusion and designed as a simplified version of MAFUNet. The main contribution of this work is combining the efficient encoder structure of Half-UNet with advanced fusion and attention mechanisms. Half-MAFUNet integrates Hierarchy Aware Mamba (HAM) for global feature modelling, Multi-Scale Adaptive Fusion (MAF) to combine global and local information, and two attention modules, Adaptive Channel Attention (ACA) and Adaptive Spatial Attention (ASA), to refine skip connections. In addition, this model incorporates Channel Atrous Spatial Pyramid Pooling (CASPP) to capture multi-scale receptive fields efficiently without increasing computational cost. Together, these components create a compact architecture that maintains strong representational power. The model is trained and evaluated on three public datasets: CVC-ClinicDB for colorectal polyp segmentation, BUSI for breast tumor segmentation, and ISIC-2018 for skin lesion segmentation. All images are resized to 256×256 pixels and processed using geometric and intensity-based augmentations. Half-MAFUNet achieves competitive performance, obtaining mean IoU around 84 85% and Dice/F1-Score around 90 92% across datasets, while using significantly fewer parameters and GFLOPs compared to U-Net, Att-UNet, UNeXt, MALUNet, LightM-UNet, VM-UNet, and UD-Mamba. These results show that Half-MAFUNet provides accurate and efficient medical image segmentation, making it suitable for real-world deployment on devices with limited computational resources.

Keywords **Medical images segmentation; Deep learning; U-Net; Efficient Model**

I. Introduction

Medical image segmentation is a crucial stage in computer-aided diagnosis and therapy. It enables clinicians to distinguish lesion areas from healthy tissue and to more clearly observe the position, size, and morphology of abnormalities. High-quality segmentation facilitates early disease detection, objective clinical assessment, and more effective treatment planning for various organs, including the colon, breast, and skin [1], [2]. In recent years, convolutional neural networks (CNNs) have shown strong performance in medical image segmentation [3], [4]. Many methods are developed based on the U-Net architecture, such as U-Net++, Attention U-Net, and other variants [5], [6]. These models achieve strong Dice and IoU performance but typically require a large number of parameters and incur substantial computational overhead. Moreover, because convolutional operations are inherently limited to local receptive fields, CNN-based architectures still struggle to capture long-range dependencies and comprehensive global context. Some works try to improve this limitation by using attention mechanisms or large-kernel convolutions [7], but these methods often make the network even heavier and more challenging to deploy in real systems. Vision Transformer-based methods and other self-attention models have also been applied to medical image segmentation [8]. They can capture global relationships among image tokens, but they typically require substantial memory and powerful GPU resources [9]. This presents a challenge for hospitals

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

or edge devices with limited hardware. Recently, selective state space models (SSMs), especially Mamba [10], have attracted attention as a new sequence modelling approach. Mamba employs a selective state-scanning approach that enables efficient modelling of long sequences while using relatively few parameters. Its vision variants, Vision Mamba and VMamba, extend this architecture to computer vision tasks, maintaining linear complexity with respect to input size and memory usage [11]. Several models, such as VM-Unet [12], HC-Mamba [13], LightM-Unet [14], Ultralight VM-Unet [15], and Polyp-Mamba [16] combine CNN and Mamba to integrate local texture and global context for segmentation.

MAFUNet is one of the latest Mamba-based architectures for medical image segmentation [17]. It introduces HAM and MAF modules, together with ACA and ASA, to improve cross-level feature interaction and multi-scale representation [18]. MAFUNet achieves high segmentation performance with fewer parameters than many prior CNN and Transformer models. However, MAFUNet still uses a relatively deep symmetric decoder and moderate model size. As a result, the computational cost and memory usage remain high for devices with limited resources and for applications requiring fast inference on multi-organ datasets. This finding indicates that a gap remains between segmentation accuracy and model efficiency in current Mamba-based medical image segmentation methods.

Accordingly, this work introduces Half-MAFUNet, a lightweight medical image segmentation architecture designed to reduce parameter count and computational cost while maintaining competitive accuracy, making it suitable for deployment on resource-constrained clinical hardware where heavier Mamba-based models are impractical. Half-MAFUNet leverages the asymmetric structure of Half-UNet, in which the decoder is intentionally simplified by using a lightweight decoding pathway rather than a fully symmetric decoder, thereby reducing the parameter count while preserving strong feature representations. The network integrates several existing modules, incorporating HAM [17] and MAF [17] to jointly exploit global context and local details, applying adaptive channel and spatial attention (ACA and ASA) [17] to selectively refine skip-connected features that exhibit multi-scale and heterogeneous lesion characteristics, and to adopt CASPP [19] in the bottleneck to efficiently enlarge the receptive field through dilated convolutions without introducing substantial additional parameters or computational overhead. The goal of this study is to build a segmentation model that maintains high accuracy with substantially lower computational cost and model size, making it more suitable for deployment in real clinical settings and on resource-limited devices [17], [20].

The main contributions of this paper can be summarized in four aspects. First, we propose Half-MAFUNet, a lightweight extension of MAFUNet that combines a Half-UNet-based architecture with Mamba and adaptive attention modules for efficient medical image segmentation. Second, we design an effective integration of HAM, MAF, CASPP, ACA, and ASA to enhance multi-scale global-local feature fusion and improve channel- and spatial-feature selection in skip connections. Third, we evaluate the proposed model using the same network architecture and training protocol on three representative public datasets, namely CVC-ClinicDB [21] for colorectal polyp segmentation, BUSI [22] for breast tumor segmentation, and ISIC-2018 [23] for skin lesion segmentation, to evaluate the generalization ability of the proposed model. Finally, we provide a detailed comparison of model complexity, including parameter count and GFLOPs, and show that Half-MAFUNet achieves competitive or better segmentation performance than several CNN-, Transformer-, and Mamba-based baselines while using fewer computational resources.

The remainder of this paper is structured as follows. Section 2 reviews CNN, Transformer, and Mamba based medical image segmentation methods. Section 3 describes the proposed Half-MAFUNet and the experimental setup (datasets, preprocessing, and training). Section 4 presents the results and ablation studies, followed by a discussion. Section 5 concludes the work and outlines future research directions.

II. Method

The methodology adopted in this study is illustrated in Fig. 1. The workflow is organized into five primary phases: (1) Dataset Preprocessing, (2) Model Architecture Design, (3) Hyperparameter Tuning, (4) Model Training, and (5) Model Evaluation.

A. Dataset Preprocessing

In this study, the CVC-ClinicDB, BUSI, and ISIC-2018 datasets are utilized. The CVC-ClinicDB dataset [21] consists of 612 polyp images, each provided with a corresponding ground-truth mask at the original resolution of 288 × 368 pixels. BUSI [22] comprises 647 breast ultrasound images with corresponding masks and varying resolutions. ISIC-2018 [23] provides 2,594 dermoscopic images with lesion masks, also at varying resolutions. The first step was resizing all images and masks to 256 × 256 pixels. The datasets were subsequently split into training, validation, and test subsets in an 80%-10%-10% proportion, where the validation set was used during training to monitor convergence, adjust the learning rate, and prevent overfitting. The detailed number of dataset splits is

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

shown in Table 1. Data augmentation was then applied to the training set to enhance data diversity and mitigate overfitting.

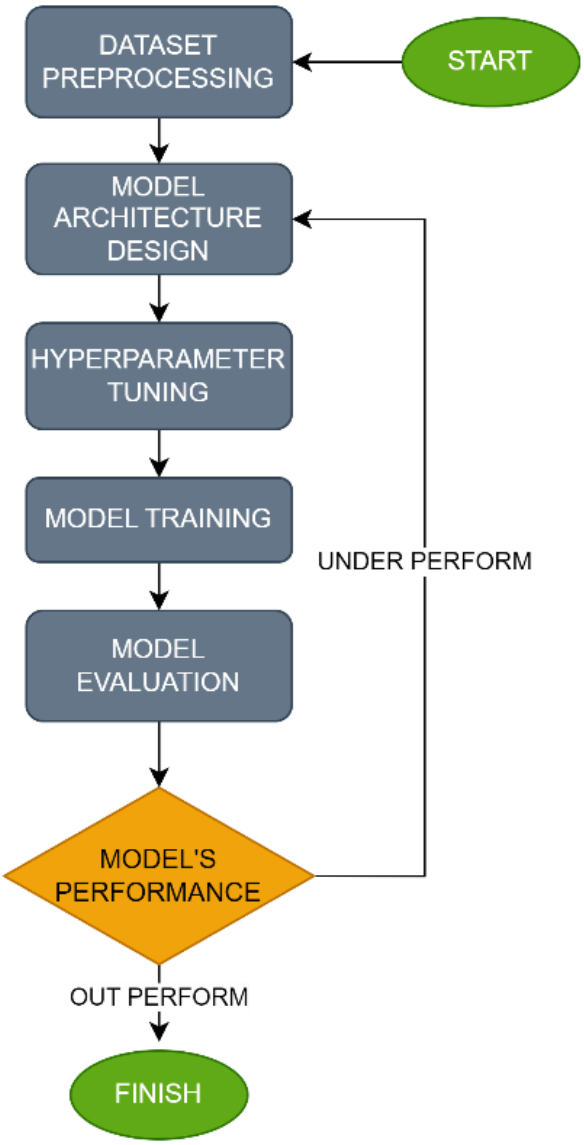


Fig. 1. Stages of Research Methods

Table 1. Number of dataset splits

Dataset	Train Set	Train Set (Augmented)	Validation Set	Test Set
CVC-ClinicDB	490	2450	61	61
BUSI	571	2585	65	65
ISIC2018	2075	4150	259	260

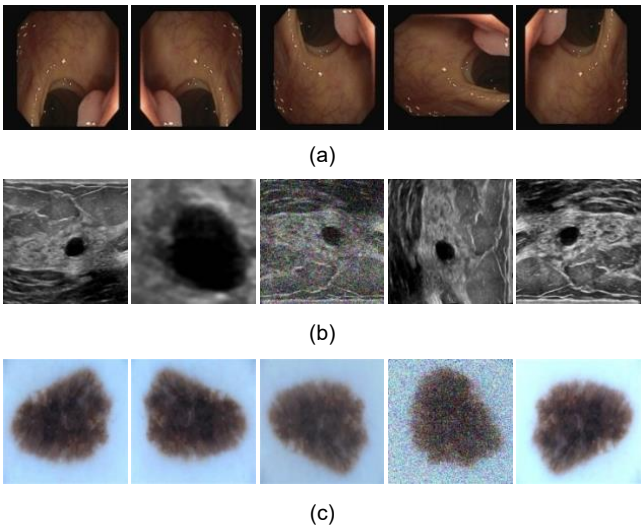


Fig. 2. Augmentation result of dataset (a) CVC-ClinicDB, (b) BUSI, and (c) ISIC-2018

Baseline augmentations comprised simple geometric transforms: horizontal flip, vertical flip, and rotations. In addition to basic geometric augmentations, dataset-specific augmentations were applied only to the training sets of BUSI and ISIC-2018 to improve robustness against modality-specific variations while avoiding information leakage. For BUSI, a pool of augmentations, including Gaussian blur, Gaussian noise, elastic transform, CLAHE, lesion-aware crop (LAC), and grid distortion, was defined to reflect common artifacts and intensity inconsistencies in ultrasound imaging. To balance data diversity and computational efficiency, each augmented sample was generated by randomly combining two to three transformations from this pool (e.g., horizontal flip with LAC and Gaussian blur), producing a limited number of representative variants per image. Similarly, for ISIC-2018, random brightness, random contrast, Gaussian noise, and grid distortion were combined in the same manner to model illumination variability and acquisition differences in dermoscopic images. The examples shown in Fig. 2 illustrate representative combinations rather than an exhaustive list, and this strategy enhances intra-dataset generalization while avoiding excessive computational overhead and overfitting. Illustrative examples of the original and augmented images are provided in Fig. 2.

B. Model Architecture Design

The overall architecture of HALF-MAFUNet is shown in Fig. 3. The network follows a five-level U-shaped topology with a Half-U-Net backbone: the encoder extracts multi-scale features, while the decoder is deliberately lightweight to reduce parameters and memory. In line with half-decoder principles, the number of filters is unified across depth levels, and feature fusion in the decoder relies primarily on

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

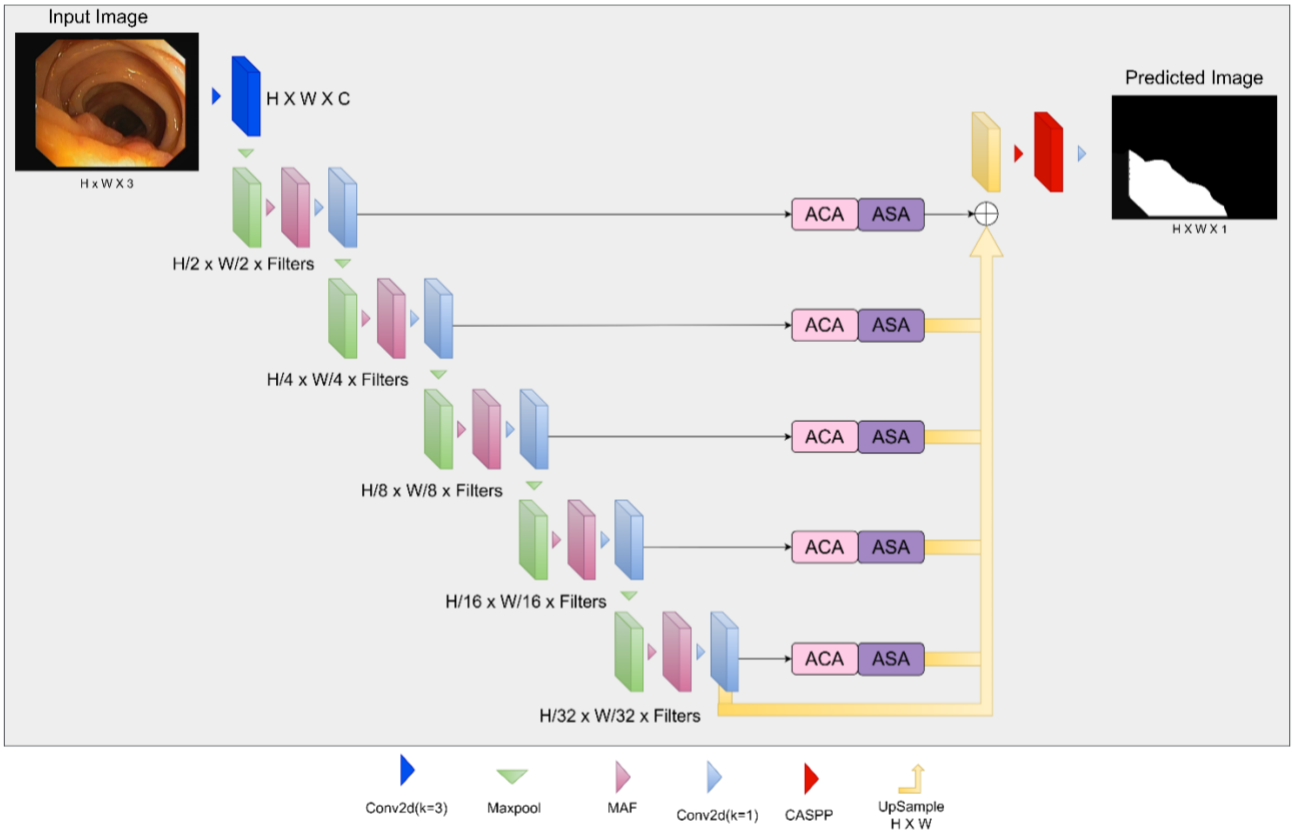


Fig. 3. Architecture of Half-MAFUNet

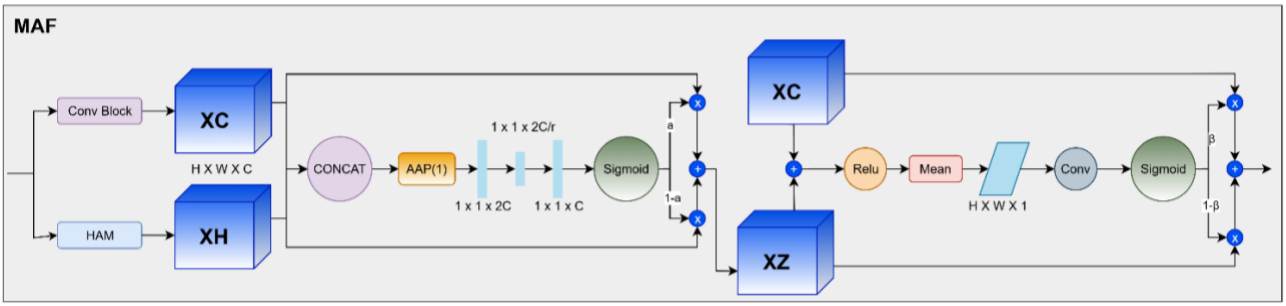


Fig. 4. Multi-scale Adaptive Fusion

element-wise addition rather than heavy concatenation, which keeps computation efficient and the design simple. At each encoder stage, features are first processed by the proposed MAF block and then passed to the next stage as well as to the skip pathway. ACA and ASA further refine the skip connections before lightweight upsampling reconstructs the final mask at the decoder side. A CASPP block is placed at the bottleneck to strengthen the multi-scale context [17].

The MAF block, illustrated in Fig. 4, adaptively fuses local and global cues. The input feature map is split into two paths: a convolutional branch X_C that focuses on texture and boundary detail, and a HAM branch X_H that encodes global structure, which are respectively obtained as $X_C = \mathcal{D}(X)$ and $X_H = \mathcal{H}(X)$ (Eqs. (1)-(2)). A channel attention weight α is produced via global average pooling followed by two fully connected layers and a sigmoid activation, thereby dynamically balancing the contributions of the convolutional and HAM branches into hybrid features (Eq. (3)). To further enhance spatial selectivity, the hybrid features are integrated with convolutional features and refined using a spatial attention mechanism. The final fusion is performed in a spatially adaptive manner, such that regions dominated by spatial attention preserve convolutional details, whereas complementary regions draw from the channel-weighted global-local

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

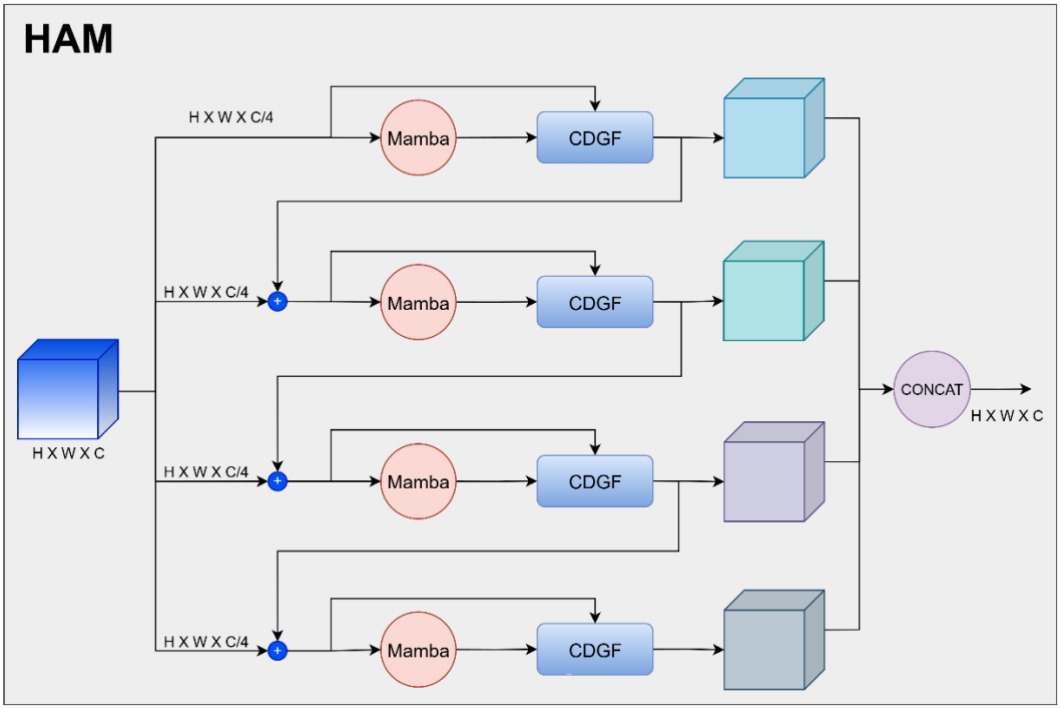


Fig. 5. Hierarchy Aware Mamba

representation, resulting in the output formulation given in Eq. (4). This cascaded channel-spatial fusion improves the complementarity between local textures and global context with minimal computational overhead [17], [24].

$$X_C = \mathcal{D}(X) \quad (1)$$

$$X_H = \mathcal{H}(X) \quad (2)$$

$$\alpha = \sigma\left(W_2\left(W_1\left(E_{(H,W)}[X_C \parallel X_H]\right)\right)\right) \quad (3)$$

$$Output = \beta \otimes X_C + (1-\beta) \otimes (\alpha \otimes X_C + (1-\alpha) \otimes X_H) \quad (4)$$

Here, X denotes the input feature map, $\mathcal{D}(\cdot)$ represents the convolutional branch that extracts local texture information, and $\mathcal{H}$ denotes the hierarchy-aware Mamba (HAM) module used for global context modelling. X_C and X_H are the local and global s representations, respectively. The operator $E_{(H,W)}$ denotes spatial average pooling, while W_1 and W_2 are fully connected layers used to generate the channel attention weight α, with $\sigma(\cdot)$ indicating the sigmoid activation function. β denotes the spatial attention weight obtained from the refined hybrid features. The symbols $\parallel$ and $\otimes$ represent channel-wise concatenation and element-wise multiplication, respectively. The output corresponds to the spatially adaptive fusion of convolutional and hybrid global-local features.

Hierarchy-Aware Mamba (HAM), illustrated in Fig. 5, is designed to enhance feature representation by jointly modelling hierarchical local interactions and long-range dependencies with low computational overhead. Given an input feature tensor $X \in \mathbb{R}^{B\times C\times H\times W}$, HAM first uniformly partitions the channel dimension into Ssub-features $\{X_i\}_{i=1}^{S}$, where each sub-feature has dimensions $X_i \in \mathbb{R}^{B\times \frac{C}{S}\times H\times W}$. A hierarchical processing flow is then constructed: at level i, the current sub-feature is combined with the output from the previous level to preserve residual information. Each hierarchical feature is processed by a Mamba block, which captures long-range dependencies and produces a global feature representation $G_i = \mathcal{M}_i(T_i)$(Eq. (5)), followed by a Channel Dynamic Gating Fusion (CDGF) unit to adaptively recalibrate channel responses [17]. To generate adaptive gating weights, global average pooling is applied to both the global feature $G_i$and the corresponding hierarchical input, and the resulting descriptors are passed through a sigmoid activation to obtain the gating vector θ, as defined in Eq. (6). The hierarchically fused feature at each level is then computed by combining the gated global feature and the previous-level output, yielding X_i' as formulated in Eq. (7). This hierarchical residual mechanism allows the local branch to preserve fine-grained spatial details while inheriting contextual information from upper layers. Finally, the gated sub-features from all levels are concatenated along the channel dimension to form the HAM output, which is forwarded to subsequent modules, as expressed in Eq. (8) [17].

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

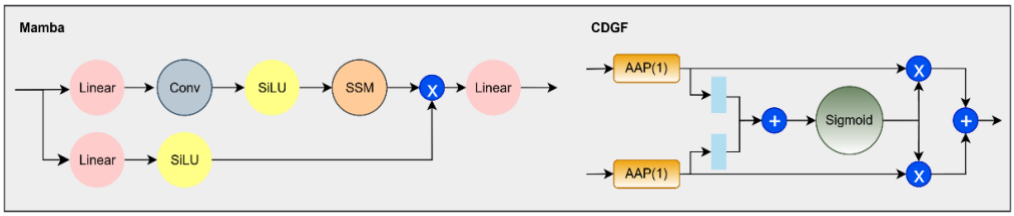


Fig. 6. Architecture of Mamba and CDGF

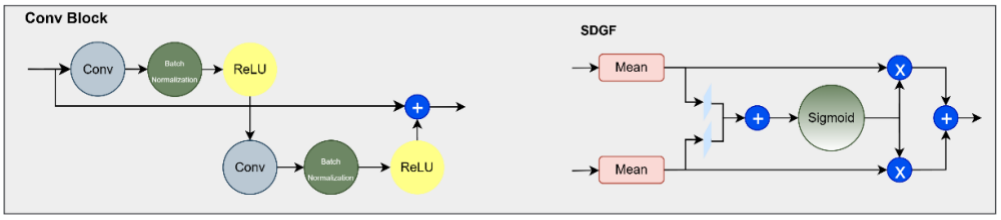


Fig. 7. Architecture of Conv block and SDGF

$$Gi = Mi(Ti) \in R^{B \times H \times W \times \frac{C}{S}} \quad (5)$$

$$\theta = \begin{cases} \sigma\left(\text{AvgPool}_{1\times1}(G_1) + \text{AvgPool}_{1\times1}(X_1)\right), & i = 1, \\ \sigma\left(\text{AvgPool}_{1\times1}(G_i) + \text{AvgPool}_{1\times1}(X_i')\right), & i > 1, \end{cases} \quad (6)$$

$$X_i' = G_i \otimes \theta + X_i^{-1'} \otimes \theta \in R^{B \times \frac{C}{S} \times H \times W} \quad (7)$$

$$Output = C(\{X_i'\}_i^{=1S}) \in R^{B \times C \times H \times W} \quad (8)$$

where $\mathcal{M}_i(\cdot)$ denotes the Mamba block at level i, T_i represents the serialized feature sequence, and G_i is the corresponding global feature output. $\sigma(\cdot)$ denotes the sigmoid activation function, AvgPool $_{1\times1}(\cdot)$ is global average pooling, $\otimes$ indicates element-wise multiplication, and X_i' represents the hierarchically fused feature at the level i. The operator $\mathcal{C}$ denotes channel-wise concatenation of all gated sub-features to form the final HAM output.

Inside HAM, illustrated in Fig. 6, the Mamba block acts as a selective state-space model that scans the serialized feature sequence. It uses content-aware gates to control how much past information is kept or updated at each step, enabling efficient modelling of long-range dependencies with linear complexity in sequence length. This mechanism is realized through a dual-collaborative architecture consisting of a main branch and an auxiliary branch that process the input sequence in parallel. The main branch captures long-range dependencies by applying a linear projection followed by convolution, SiLU activation, and a selective state-space model, producing $\text{Branch}_1(X)$ as formulated in Eq. (9). In parallel, the auxiliary branch applies a lightweight linear transformation with SiLU activation to preserve the basic feature representation, yielding $\text{Branch}_2(X)$ as defined in Eq. (10). The outputs of the two branches are then fused through element-wise multiplication and mapped by a linear layer to obtain the final Mamba output, as expressed in Eq. (11). Compared with self-attention, this dual-branch selective fusion allows Mamba to provide global receptive fields with fewer parameters and more hardware-friendly computation, making it well suited for lightweight medical image segmentation [17].

$$\text{Branch}_1(X) = \text{SSM}\left(\text{SiLU}\left(\text{Conv}\left(\text{Linear}(X)\right)\right)\right) \quad (9)$$

$$\text{Branch}_2(X) = \text{SiLU}\left(\text{Linear}(X)\right) \quad (10)$$

$$\text{Mamba}(X) = \text{Linear}\left(\text{Branch}_1(X) \otimes \text{Branch}_2(X)\right) \quad (11)$$

Throughout HALF-MAFUNet, convolutional blocks serve as basic local feature extractors, as illustrated in Fig. 7. In the encoder and decoder, each block typically consists of convolution, batch normalization, and ReLU activation. Standard or depthwise separable convolutions are employed at each stage to balance expressive power and efficiency. These blocks provide stable local feature representations that are later enhanced by HAM, MAF, ACA, ASA, and CASPP, forming the backbone of the network's hierarchical feature hierarchy [17]. Channel Dynamic Gating Fusion (CDGF), illustrated in Fig. 6 operates on the outputs of each HAM level. Given the global gating vector θ, CDGF scales the channels of each sub-feature and combines them in a level-aware manner. This mechanism allows the model to emphasize informative levels and suppress redundant responses when aggregating multi-level features. By integrating CDGF into HAM, HALF-MAFUNet can adaptively balance contributions from different depths, improving global-local interaction without significantly increasing computation [17]. Spatial Dynamic Gating Fusion (SDGF), illustrated in Fig. 7, combines an original feature map with its spatially attended counterpart. It first extracts pooled statistics (for example, global average and max pooling) from both inputs to generate a spatial gate. This gate determines the extent to which

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

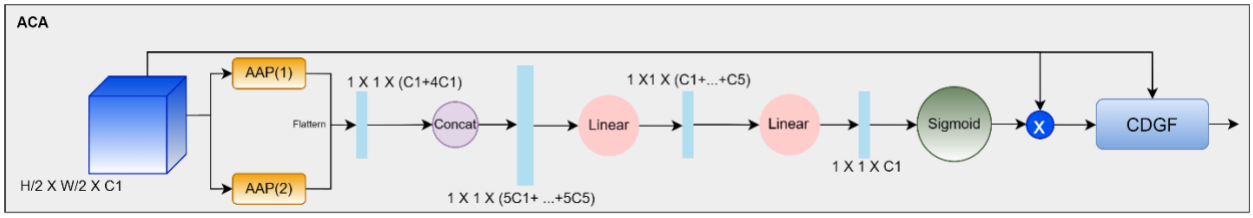


Fig. 8. Architecture of adaptive channel attention

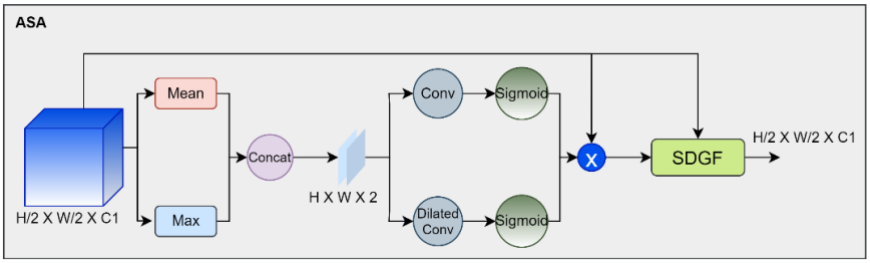


Fig. 9. Architecture of adaptive spatial attention

information should be drawn from the original feature and from the refined one at each spatial location. In this way, SDGF provides a flexible mechanism for suppressing noise while preserving important spatial structures, especially when applied after spatial attention modules [17].

Adaptive Channel Attention (ACA) on skips, as illustrated in Fig. 8, improves the representational quality of features transmitted through skip connections by dynamically recalibrating channel responses. For each level, ACA aggregates multi-level global context using adaptive average pooling at two spatial scales (e.g., $1 \times 1$and 2×2) to capture contextual information at different granularities, and the resulting pooled features are flattened and concatenated along the channel dimension to form the pooled descriptor F_k^{pool}, as defined in Eq. (12). The pooled descriptors from all levels are then concatenated to construct a global feature representation F^{global} (Eq. (13)), which is subsequently compressed through a linear layer to suppress redundant information, yielding the reduced feature F^{reduced} (Eq. (14)). Based on this reduced representation, level-specific fully connected layers followed by sigmoid activation generate adaptive channel attention weights α_k for each level, as formulated in Eq. (15). These weights are expanded and multiplied with the corresponding input feature maps to recalibrate informative channels, producing the channel-refined features T_k' (Eq. (16)). Finally, a light spatial gating term θ, computed from the combined global pooling of the recalibrated and original features, is applied to stabilize the feature responses, and the refined skip feature delivered to the decoder is obtained as expressed in Eqs. (17)-(18) [17].

$$F_k^{\text{pool}} = \text{Flatten}\left(\text{AvgPool}_{1\times1}(T_k)\right) \oplus \text{Flatten}\left(\text{AvgPool}_{2\times2}(T_k)\right) \quad (12)$$

$$F^{\text{global}} = \|_{k=1}^{5} F_k^{\text{pool}} \quad (13)$$

$$F^{\text{reduced}} = W_r\left(F^{\text{global}}\right) \quad (14)$$

$$\alpha_k = \sigma\left(W_{\alpha_k}\left(F^{\text{reduced}}\right)\right) \quad (15)$$

$$T_k' = \alpha_k \otimes T_k \quad (16)$$

$$\theta = \sigma\left(\text{AvgPool}_{1\times1}(T_k') + \text{AvgPool}_{1\times1}(T_k)\right) \quad (17)$$

$$T_k^{\text{output}} = T_k' \otimes \theta + T_k \otimes \theta \quad (18)$$

Here, $T_k \in \mathbb{R}^{B\times C_k\times H_k\times W_k}$ denotes the input feature map at level k of the skip connection. $\text{AvgPool}_{1\times1}(\cdot)$ and $\text{AvgPool}_{2\times2}(\cdot)$ represent adaptive average pooling operations at different spatial scales, while flatten converts pooled features into channel descriptors. The operator $\oplus$ denotes channel-wise concatenation, and $\|_{k=1}^{5}(\cdot)$ indicates cross-level concatenation of pooled features. W_r is a linear transformation used for dimensionality reduction, and W_{α_k} denotes level-specific fully connected layers that generate the channel attention weights α_k. The sigmoid function σ ensures that attention weights are normalized to the range [0, 1]. The symbol $\otimes$ denotes element-wise multiplication. The spatial gating term θstabilizes feature recalibration by jointly considering the original and channel-refined features, and T_k^{output} represents the final ACA-refined skip feature delivered to the decoder.

Adaptive Spatial Attention (ASA) on skips, as shown in Fig. 9, complements ACA by emphasizing where to focus in the spatial domain. Given a feature map F_i, ASA first computes global average-pooled and max-

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

pooled summaries along the channel dimension to highlight informative spatial cues. These pooled features are then processed by two convolutional paths to extract spatial dependencies: a dilated convolution path with a kernel size of 7×7 and dilation rate 3 to enlarge the receptive field, and a standard 3×3 convolution path to capture fine-grained local structures. The outputs of these two paths are fused to generate a spatial attention map A_i^s, as formulated in Eq. (19). The spatial attention map is subsequently applied to the input feature map to obtain the spatially refined feature F_i' (Eq. (20)). To further stabilize feature fusion, a spatial gating term θ_i is computed from the combined global average pooling of the original and spatially refined features, as defined in Eq. (21). Finally, Spatial Dynamic Gating Fusion (SDGF) is employed to adaptively combine the original feature F_i and the spatially attended feature F_i' using the gating term, yielding the optimized skip feature for decoding, as expressed in Eq. (22) [17].

$$A_i^s = \sigma\left(C_{d=3}^{7\times7}\left(\left[F_{\text{avg}}; F_{\text{max}}\right]\right) + C^{3\times3}\left(\left[F_{\text{avg}}; F_{\text{max}}\right]\right)\right) \quad (19)$$

$$F_i' = A_i^s \otimes F_i \quad (20)$$

$$\theta_i = \sigma\left(F_{\text{avg}}(F_i) + F_{\text{avg}}(F_i')\right) \quad (21)$$

$$F_i^{\text{output}} = F_i \otimes \theta_i + F_i' \otimes \theta_i \quad (22)$$

Here, $F_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ denotes the input skip feature at level i. $F_{\text{avg}}$and F_{max} represent global average pooling and max pooling operations along the channel dimension, respectively. $C_{d=3}^{7\times7}$ and $C^{3\times3}$ denote dilated and standard convolution operations used to capture large-scale and local spatial dependencies. The sigmoid function σ produces the spatial attention map A_i^s, which modulates the input feature map to obtain the refined feature F_i'. The spatial gating term θ_i is computed from pooled statistics of both original and refined features to stabilize feature fusion. The symbol $\otimes$ denotes element-wise multiplication, and F_i^{output} represents the final ASA-refined skip feature delivered to the decoder.

The Channel Atrous Spatial Pyramid Pooling (CASPP) block, illustrated in Fig. 10, aims to capture rich multi-scale context without a substantial increase in computational cost. Starting from an input feature map with C_{in}channels, CASPP builds several parallel branches with different dilation rates (e.g., 1, 6, 12, and 18). For rate $r = 1$, a 1×1 convolution is used to preserve local detail, while for $r > 1$, 3×3 dilated convolutions with padding $p = r$ expand the receptive field and capture larger structural patterns. Each branch produces C_{br} feature channels, followed by batch normalization and ReLU activation to stabilize the responses. The outputs of all branches are concatenated along the channel dimension and then compressed by a 1×1 projection layer with batch normalization and ReLU, yielding an output with C_{out} channels. Placed at the bottleneck of HALF-MAFUNet, CASPP provides a compact yet expressive representation that encodes lesion context at multiple scales, which is especially helpful for handling objects with highly variable sizes and shapes [19], [24], [25], [26].

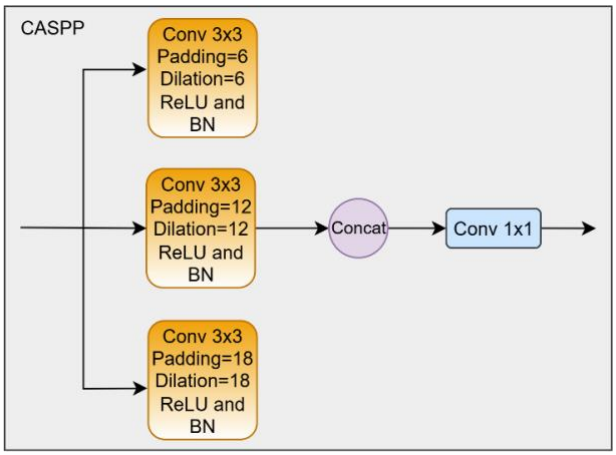


Fig. 10. Architecture of CASPP

C. Hyperparameter Tuning

Next, hyperparameters are tuned to strike an appropriate trade-off between segmentation accuracy and computational efficiency. The initial hyperparameter choices and explored ranges are determined by common practices in lightweight medical image segmentation and prior related work, to ensure stable convergence while maintaining efficiency before detailed tuning is conducted. Hyperparameter selection directly affects both the convergence speed and the model's ability to generalize to unseen test data. As summarized in Table 2, the initial configuration in this study uses a batch size of 8 and up to 100 training epochs to balance training stability and memory constraints, the Adam optimizer for its reliable convergence behavior, and an initial learning rate of 0.001, which is widely adopted in segmentation tasks. The learning rate is reduced by a factor of 10 if the validation loss fails to improve for 10 consecutive epochs, thereby stabilizing the training process. Binary Cross Entropy (BCE) is adopted as the loss function due to its effectiveness in binary segmentation, and the network depth is fixed at five levels, consistent with the Half-U-Net design to preserve sufficient representational capacity. In addition, the attention ratio is set to 8, and the number of filters is set to 64 channels at each stage of the encoder-decoder, as a compromise between model compactness and segmentation performance. These settings are then used as the default configuration for all subsequent training on the three datasets.

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Table 2. Initial Hyperparameters

Hyperparameter	Value
Batch Size	8
Epoch	100
Optimizer	Adam
Learning Rate	0.001 (reduced to one-tenth if the validation loss fails to improve for 10 epochs.)
Loss Function	Binary Cross Entropy
Depth	5
Attention Ratio	8
Filters	64

D. Model Training

Once the optimal hyperparameters are determined, the proposed HALF-MAFUNet is trained using the training portion of each dataset. The selected configuration strongly influences the final performance, with 256×256 RGB medical images as input and a binary. Segmentation mask as the output. Optimization employs AdamW with a ReduceOnPlateau learning rate schedule and Binary Cross-Entropy loss. Model training is implemented in PyTorch on a high-performance computer (HPC) system with GPU acceleration using an NVIDIA RTX A4000.

E. Model Evaluation

Model performance is assessed using several quantitative metrics to evaluate the quality of the segmentation results. In this study, we use mean Intersection over Union (mIoU), Dice coefficient (F1-Score), Accuracy (Acc), Sensitivity (Sen), and Specificity (Spe). These metrics are widely used in medical image segmentation because they measure both the overlap between the predicted mask and the ground truth and the accuracy with which the model identifies positive and negative pixels. The formulas for each metric are shown in Eqs. (23)-(27).

$$mIoU = \frac{TP}{TP + FP + FN} \tag{23}$$

$$Dsc = \frac{2TP}{2TP + FP + FN} \tag{24}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

$$Sen = \frac{TP}{TP + FN} \tag{26}$$

$$Spe = \frac{TN}{TN + FP} \tag{27}$$

In this work, TP (true positive) denotes the count of lesion pixels that the model correctly classifies as a lesion, while TN (true negative) represents the count of background pixels correctly identified as background. FP (false positive) is the number of background pixels that are mistakenly labeled as a lesion, and FN (false negative) is the number of lesion pixels that are wrongly labeled as background. In our experiments, these metrics are first computed for each test image and then averaged across all test images in the dataset.

III. Result

A. Hyperparameter Tuning

In this study, hyperparameter tuning was performed by varying several key components of the HALF-MAFUNet model, including the optimizer, attention ratio, loss

Table 3. Hyperparameter tuning test results

Hyperparameter		Params(M)	FLOPs(G)	mIoU(%)	F1- Score(%)	ACC(%)	SEN(%)	SPE(%)
Optimizer	Adam	0.5373	7.944	86.37	91.70	98.76	91.66	99.49
	AdamW	**0.5373**	**7.944**	**88.68**	**93.38**	**98.94**	**96.75**	**99.17**
Attention Ratio	4	0.5449	7.944	86.36	91.72	98.82	93.61	99.36
	8	**0.5373**	**7.944**	**88.68**	**93.38**	**98.94**	**96.75**	**99.17**
	16	0.5334	7.944	86.29	91.60	98.80	93.97	99.30
Loss Function	BCE	0.5373	7.944	86.05	91.45	98.75	92.76	99.37
	Dice Loss	0.5373	7.944	86.66	91.75	98.70	93.14	99.28
	BCE + Dice Loss	**0.5373**	**7.944**	**88.68**	**93.38**	**98.94**	**96.75**	**99.17**
Depth	3	0.3512	7.380	81.67	88.53	98.04	88.07	99.08
	4	0.4442	7.670	87.03	92.11	98.85	93.19	99.44
	5	**0.5373**	**7.944**	**88.68**	**93.38**	**98.94**	**96.75**	**99.17**
Filters	16	0.0390	0.682	74.16	82.47	97.48	81.53	99.13
	32	0.1368	2.049	81.61	88.48	98.37	89.75	99.26
	64	**0.5373**	**7.944**	**88.68**	**93.38**	**98.94**	**96.75**	**99.17**

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Table 4. Ablation experiment results

Dataset	Method	Params(M)	FLOPs(G)	mIoU(%)	F1- Score(%)	ACC(%)	SEN(%)	SPE(%)
CVC-ClinicDB	Ablation 1	0.5371	7.943	83.51	89.84	98.48	90.28	99.33
	Ablation 2	0.5296	7.944	87.66	92.84	99.00	94.53	99.46
	Ablation 3	0.5373	7.943	85.81	91.20	98.73	92.02	99.43
	Ablation 4	**0.5373**	**7.944**	**88.68**	**93.38**	**98.94**	**96.75**	**99.17**
BUSI	Ablation 1	0.5371	7.943	81.04	88.57	98.11	90.17	98.91
	Ablation 2	0.5296	7.944	81.01	88.64	98.20	91.87	98.84
	Ablation 3	0.5373	7.943	80.63	88.38	98.14	90.72	98.89
	Ablation 4	**0.5373**	**7.944**	**83.83**	**90.50**	**98.37**	**92.42**	**98.97**
ISIC-2018	Ablation 1	0.5371	7.943	79.68	87.34	94.51	84.06	97.47
	Ablation 2	0.5296	7.944	79.99	87.49	94.82	81.46	98.60
	Ablation 3	0.5373	7.943	78.31	86.10	94.10	78.59	98.48
	Ablation 4	**0.5373**	**7.944**	**80.15**	**87.68**	**94.71**	**83.12**	**97.99**

function, network depth, and number of filters. To save time, all tuning experiments were only run on the CVC-ClinicDB dataset. The process was conducted step by step: the initial setting in Table 2 was updated whenever a better configuration was found. The complete tuning results are shown in Table 3.

For the optimizer, we compared Adam and AdamW while keeping other hyperparameters fixed. AdamW achieved the best overall performance, with an mIoU of 88.68%, F1-Score of 93.38%, accuracy of 98.94%, sensitivity of 96.75%, and specificity of 99.17%, outperforming Adam in all metrics except specificity, which was slightly higher for Adam. Next, we evaluated the attention ratio with values of 4, 8, and 16. An attention ratio of 8 yielded the highest scores (mIoU 88.68%, F1- Score 93.38%), while ratios 4 and 16 produced lower performance, indicating that too small or too large attention bottlenecks reduce the effectiveness best attention modules. Different loss functions were then examined, including BCE, Dice Loss, and a combined trade-off, again achieving an mIoU of 88.68% and an F1-score of 93.38%, and slightly improving sensitivity compared to using BCE or Dice alone. This suggests that combining region-based and overlap-based objectives helps the model learn more balanced foreground-background segmentation. We also analyzed the effect of network depth and filter size on performance, observing that increasing the depth from 3 to 5 levels consistently improved segmentation accuracy of the BCE + Dice Loss. The hybrid loss achieved the with a depth of 5 achieving the best results (mIoU 88.68%, F1-Score 93.38%) at a modest increase in parameters (from 0.3512M to 0.5373M) while reducing the number of filters to 16 or 32 substantially lowered parameters and GFLOPs but resulted in inferior segmentation performance; in contrast, using 64 filters provided the most favorable accuracy-efficiency trade-off, maintaining an acceptable computational cost relative to existing CNN- and Mamba-based models while achieving competitive segmentation performance, and was therefore selected as the final configuration for all experiments.

B. Model Ablation

To comprehensively evaluate the contribution of the attention gate design in the proposed HALF-MAFUNet, several ablation experiments were carried out on three public datasets: CVC-ClinicDB, BUSI, and ISIC-2018. In all settings, the backbone, HAM, MAF, and CASPP blocks were intentionally kept unchanged to maintain a fixed architectural backbone and isolate the impact of different attention gate designs on the skip connections, ensuring that the observed performance differences originate solely from the attention mechanisms rather than changes in the core feature extraction or multi-scale context modeling modules. Four variants were tested: (1) Ablation 1, HALF-MAFUNet with an attention gate that uses only ACA (without ASA); (2) Ablation 2, HALF-MAFUNet with an attention gate that uses only ASA (without ACA); (3) Ablation 3, HALF-MAFUNet with an attention gate that applies ASA first and then ACA; and (4) Ablation 4, HALF-MAFUNet with an attention gate that applies ACA first and then ASA, which is our proposed method. All variants exhibit similar computational complexity, with parameters of approximately 0.53M and FLOPs of approximately 7.94G, so performance differences primarily arise from the attention design.

As summarized in Table 4, the proposed attention order (Ablation 4) consistently achieves the best balance

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

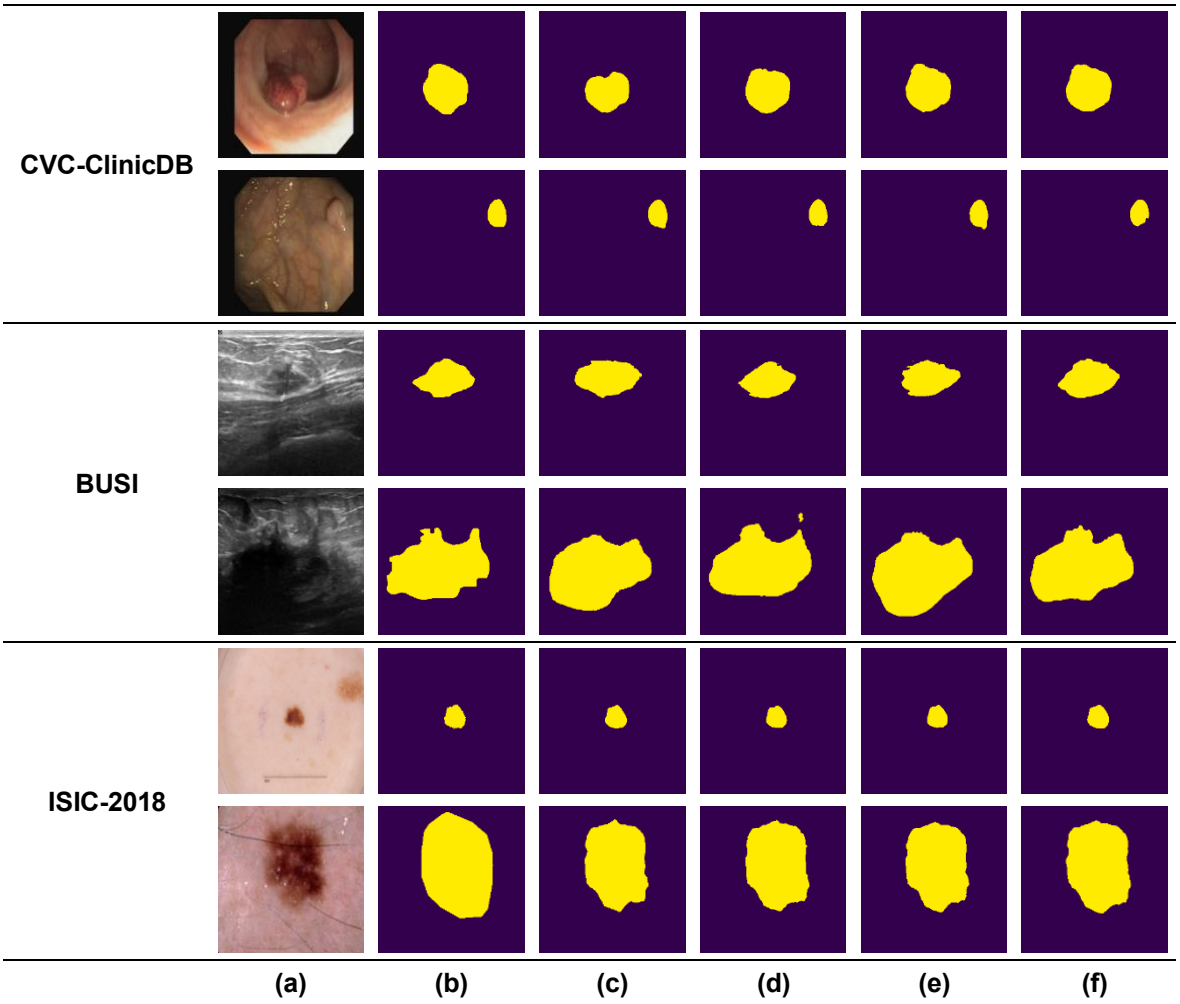


Fig. 11. Qualitative comparison of model ablation: (a) input image, (b) ground truth, (c) ablation 1, (d) ablation 2, (e) ablation 3, (f) ablation 4

across all metrics. On the CVC-ClinicDB dataset, Ablation 4 obtains the highest mIoU of 88.68% and F1-score of 93.38%, with accuracy of 98.94%, sensitivity of 96.75%, and specificity of 99.17%. Compared with using only ACA (Ablation 1) or only ASA (Ablation 2), the dual-module gate in Ablation 4 yields a clear improvement of approximately 5% mIoU and more than 3% F1-score over the plain ACA case. On the BUSI dataset, Ablation 4 again achieves the best results, reaching an mIoU of 83.83% and an F1-score of 90.50%, and also yields the highest sensitivity (92.42%) and specificity (98.97%) among all variants. These results show that combining channel and spatial attention in a cascaded manner is more effective than using a single attention type.

On the ISIC-2018 dataset, Ablation 4 achieves the highest mIoU and F1-score, with values of 80.15% and 87.68%, respectively. Although Ablation 2 slightly increases accuracy and specificity, the proposed configuration still provides the best overlap-based metrics, which are more important for lesion segmentation quality. Overall, the comparison indicates that both ACA and ASA are necessary, and that their application order plays a critical role: applying ACA first enables channel-wise filtering to suppress less, thereby more effectively refining lesion localization by focusing on spatially relevant regions, which ultimately yields more reliable segmentation results. Qualitative examples in Fig. 11 further confirm that Ablation 4 produces lesion masks that are more complete and closer to the ground-truth boundaries than those of other ablation variants, particularly in challenging cases with irregular shapes or low contrast.

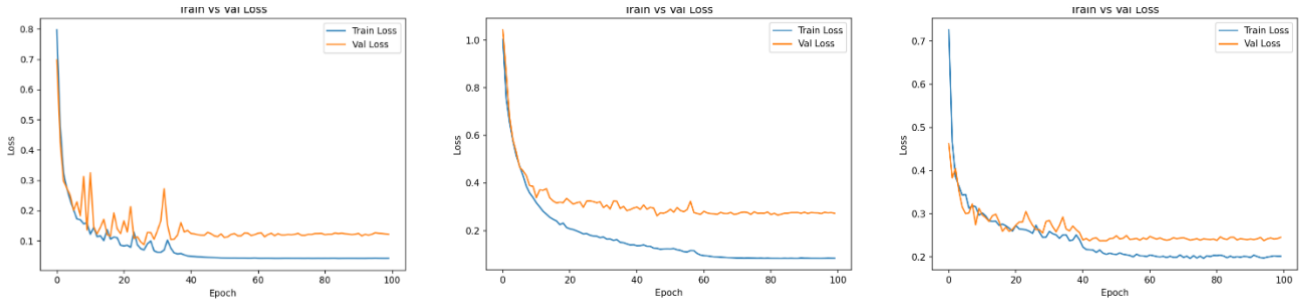
C. Experimental Result

Hyperparameter tuning and ablation experiments resulted in the optimal configuration of HALF-MAFUNet, whose quantitative performance is summarized in Table 5. With only 0.54M parameters and 7.94 GFLOPs, the model achieves mIoU/F1-Scores of 88.68%/93.38% on

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Table 5. Experiment result

Dataset	Params(M)	FLOPs(G)	mIoU(%)	F1-Score(%)	ACC(%)	SEN(%)	SPE(%)
CVC-ClinicDB	0.5373	7.944	88.68%	93.38%	98.94%	96.75%	99.17%
BUSI	0.5373	7.944	83.83%	90.50%	98.37%	92.42%	98.97%
ISIC-2018	0.5373	7.944	80.15%	87.68%	94.71%	83.12%	97.99%



(a) (b) (c)

Fig. 12. Training and validation loss on dataset: (a) CVC-ClinicDB, (b) BUSI, (c) ISIC-2018

CVC-ClinicDB, 83.83%/90.50% on BUSI, and 80.15%/87.68% on ISIC-2018, while accuracy, sensitivity, and specificity exceed 94% across all datasets. All results are obtained using predefined training, validation, and test sets without cross-validation, and the robustness of the proposed model is evidenced by consistent performance across three datasets and stable convergence behavior during training.

Table 6. Experiment Result

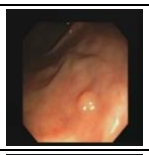
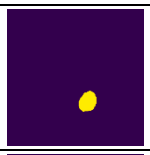
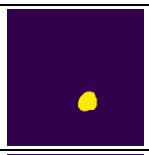
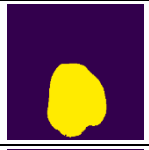
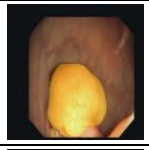
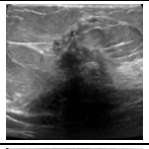
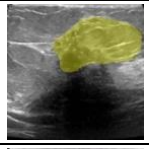
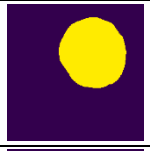
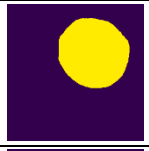
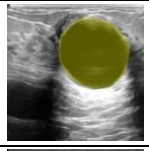
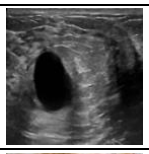
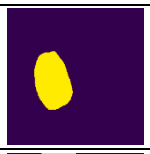
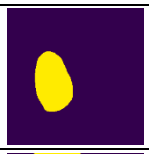
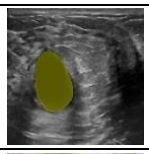
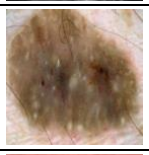
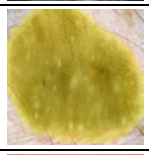
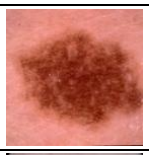
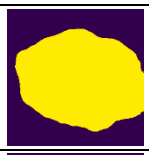
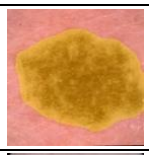
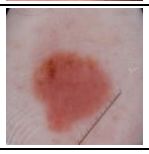
Hyperparameter	Value
Batch Size	8
Epoch	100
Optimizer	AdamW
Learning Rate	0.001 (reduced to one-tenth if the validation loss fails to improve for 10 epochs.)
Loss Function	BCE + Dice Loss
Depth	5
Attention Ratio	8
Filters	64

The best hyperparameter setting used to obtain these results is listed in Table 6. It uses a batch size of 8, 100 training epochs, the AdamW optimizer, an initial learning rate of 0.001 with step decay if the validation loss does not improve for 10 epochs, BCE + Dice Loss, a network depth of 5, an attention ratio of 8, and 64 filters. With this setup, the proposed model is trained on three datasets: CVC-ClinicDB, BUSI, and ISIC-2018. Example results for images from each dataset are presented in Table 7, which compares the ground-truth masks with the predicted masks. These examples show that the proposed method can follow lesion edges and shapes quite well, with good completeness and smooth boundaries, informative feature maps, allowing the subsequent ASA.

To further analyze the limitations of the proposed model, the challenging cases illustrated in Table 8 provide additional insight into scenarios where segmentation performance degrades. For CVC-ClinicDB, errors typically occur when polyps are very small or have low contrast with the surrounding mucosa, resulting in incomplete region coverage. In BUSI, strong speckle noise and ambiguous lesion boundaries in ultrasound images often lead to under-segmentation or boundary inaccuracies. Similarly, in ISIC-2018, lesions with irregular shapes, low color contrast, or heterogeneous textures remain challenging and can occasionally cause partial segmentation or shape distortion. As shown in Table 8, these failure patterns are closely related to modality-specific imaging characteristics rather than systematic weaknesses of the proposed architecture, highlighting directions for further improvement in handling low-contrast regions and complex lesion boundaries.

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Table 7. Experiment result

Dataset	Input Image	Ground Truth	Predicted Image	Overlay	F1- Score / IoU (%)
CVC-ClinicDB					95.05 / 90.57
					97.73 / 95.56
					97.82 / 95.74
BUSI					94.58 / 89.73
					96.57 / 93.37
					95.77 / 91.88
ISIC-2018					95.27 / 90.96
					92.61 / 85.24
					92.50 / 87.80

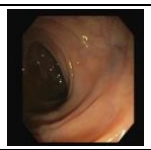
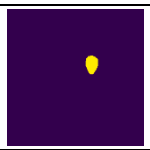
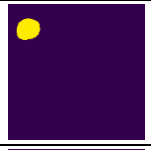
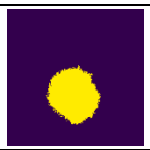
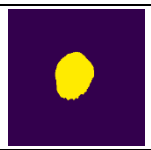
During training, we tracked both training and validation losses to assess convergence and monitor potential overfitting. As shown in Fig. 12, the loss curves gradually decreased and then stabilized as the number of epochs increased, indicating smooth convergence under the selected hyperparameters. No severe divergence or strong overfitting trend was observed, suggesting that the chosen configuration provides a good balance between model capacity and regularization across all three medical image segmentation tasks.

IV. Discussion

This study proposes HALF-MAFUNet, a lightweight medical image segmentation architecture that balances segmentation accuracy and computational efficiency across different imaging modalities. As summarized in

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Table 8. Challenging experiment result

Dataset	Input Image	Ground Truth	Predicted Image	Overlay	F1- Score / IoU (%)
CVC-ClinicDB					18.85 / 10.41
BUSI					32.76 / 19.59
ISIC-2018					49.02 / 33.27

Tables 9-11, the model achieves consistently high mIoU, F1-score, accuracy, sensitivity, and specificity on CVC-ClinicDB, BUSI, and ISIC-2018 while using only about 0.54 million parameters and 7.94 GFLOPs. The best results are obtained on CVC-ClinicDB (mIoU 88.68%, F1-Score 93.38%), followed by BUSI (mIoU 83.83%, F1-Score 90.50%) and ISIC-2018 (mIoU 80.15%, F1-Score 87.68%). Variations in performance across datasets can be attributed to the distinct characteristics of each imaging modality. CVC-ClinicDB images generally exhibit more precise lesion boundaries and relatively homogeneous backgrounds, which facilitate accurate polyp segmentation.

In contrast, BUSI ultrasound images are characterized by strong speckle noise and ambiguous lesion boundaries, which often lead to boundary inaccuracies or partial under-segmentation. Similarly, ISIC-2018 dermoscopic images present greater challenges due to substantial variations in color, texture, illumination conditions, and lesion shapes, which may result in incomplete region coverage or shape distortion. These failure patterns are primarily driven by modality-specific imaging properties rather than systematic weaknesses of the proposed model, and they highlight directions for future improvements in handling low-contrast regions and complex lesion boundaries.

Table 9 shows that on CVC-ClinicDB, HALF-MAFUNet achieves the highest mIoU and F1-score among all compared methods while using far fewer parameters than classic CNN models such as U-Net and Att-UNet, and even fewer than recent Mamba-based MAFUNet and UD-Mamba. At the same time, the model attains very high accuracy (98.94%), sensitivity (96.75%), and specificity (99.17%), indicating that it can detect most polyp pixels while keeping false positives low. This suggests that the combination of Half-UNet backbone, HAM, MAF, CASPP, and ACA ASA attention can compensate for the reduced decoder size and still outperform heavier architectures.

As shown in Table 10, BUSI, HALF-MAFUNet again achieves the best mIoU and F1-Score (83.83% and 90.50%), outperforming U-Net, UNeXt, MALUNet, LightM-UNet, and other recent architectures. Sensitivity reaches 92.42%, which is important for detecting breast tumors that often exhibit blurred boundaries on ultrasound images, while accuracy and specificity remain above 98%. Compared to other lightweight competitors with similar parameter budgets, such as MALUNet or Ultralight VM-UNet, this improvement can be attributed to the explicit modeling of global context through HAM and the adaptive global-local feature fusion enabled by MAF, which helps mitigate the strong noise and boundary ambiguity commonly present in ultrasound images. In addition, the attention gate on skip connections further refines feature selection, allowing the decoder to focus on more informative representations despite its lightweight design. These results indicate that the proposed model can generalize well from colonoscopy to ultrasound images without changing the architecture.

As shown in Table 11 for ISIC-2018, HALF-MAFUNet obtains slightly lower mIoU and F1-Score than the heavier UD-Mamba and MAFUNet models, yet it still achieves competitive performance with mIoU 80.15% and F1-Score 87.68%. While skin lesion images exhibit high variability in color, texture, and lesion shape, the combination of global context modeling, adaptive fusion, and selective attention allows HALF-MAFUNet to maintain robust performance with significantly fewer parameters. Considering that HALF-MAFUNet uses 18-60 times fewer parameters than many Transformer- and Mamba-based competitors, these results highlight a favorable trade-off between segmentation quality and

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Table 9. Quantitative comparison with previous methods on dataset CVC-ClinicDB

Model	Year	Params(M)	FLOPs(G)	mIoU(%)	F1- Score(%)	ACC(%)	SEN(%)	SPE(%)
U-Net [5]	2015	31.03	54.73	83.53	90.96	98.36	88.23	**99.42**
Att-UNet [18]	2018	34.88	66.63	84.35	91.39	98.44	89.83	99.31
UNeXt [27]	2022	1.47	0.57	70.15	81.87	96.87	77.13	98.88
MALUNet [28]	2022	0.18	0.08	74.71	85.09	97.29	82.07	98.87
UTNetV2 [6]	2022	12.80	15.50	84.98	91.79	98.48	91.34	99.20
FocalUNETR [29]	2023	26.91	16.28	82.83	90.46	98.24	90.02	99.05
LightM-UNet [14]	2024	0.19	0.66	70.94	82.61	96.71	82.81	98.17
Ultralight VM-Unet [15]	2024	**0.05**	**0.06**	73.47	84.24	97.26	80.12	99.02
VM-UNet [12]	2024	22.04	4.11	84.31	91.37	98.43	90.18	99.27
U-KAN [30]	2025	9.38	6.89	84.92	91.67	98.50	91.50	99.19
UD-Mamba [31]	2025	19.12	5.91	84.83	91.73	98.51	90.98	99.23
MAFUNet [17]	2025	9.61	7.43	85.07	91.85	98.53	89.90	99.40
Half-MAFUNet	**Proposed Model**	0.53	7.944	**88.68**	**93.38**	**98.94**	**96.75**	99.17

Table 10. Quantitative comparison with previous methods on dataset BUSI

Model	Year	Params(M)	FLOPs(G)	mIoU(%)	F1- Score(%)	ACC(%)	SEN(%)	SPE(%)
U-Net [5]	2015	31.03	54.73	66.93	79.65	96.50	81.79	97.92
Att-UNet [18]	2018	34.88	66.63	67.51	80.11	96.75	75.93	98.78
UNeXt [27]	2022	1.47	0.57	66.11	79.16	96.61	74.83	98.74
MALUNet [28]	2022	0.18	0.08	62.12	75.42	95.63	77.79	97.32
UTNetV2 [6]	2022	12.80	15.50	70.92	82.39	97.10	81.12	98.64
FocalUNETR [29]	2023	26.91	16.28	68.57	80.79	96.93	76.18	98.95
LightM-UNet [14]	2024	0.19	0.66	65.41	78.47	96.47	79.45	98.07
Ultralight VM-Unet [15]	2024	**0.05**	**0.06**	62.78	76.51	95.93	78.26	97.65
VM-UNet [12]	2024	22.04	4.11	69.29	81.36	96.81	80.30	98.44
U-KAN [30]	2025	9.38	6.89	70.40	82.04	97.10	80.42	98.70
UD-Mamba [31]	2025	19.12	5.91	71.17	82.71	97.31	81.60	98.55
MAFUNet [17]	2025	9.61	7.43	71.68	83.12	97.19	82.45	98.61
Half-MAFUNet	**Proposed Model**	0.53	7.944	**83.83**	**90.50**	**98.37**	**92.42**	**98.97**

Table 11. Quantitative comparison with previous methods on dataset ISIC-2018

Model	Year	Params(M)	FLOPs(G)	mIoU(%)	F1- Score(%)	ACC(%)	SEN(%)	SPE(%)
U-Net [5]	2015	31.03	54.73	77.86	87.55	94.05	85.86	96.69
Att-UNet [18]	2018	34.88	66.63	78.43	87.91	94.13	87.60	96.23
UNeXt [27]	2022	1.47	0.57	79.50	88.58	94.59	86.18	97.29
MALUNet [28]	2022	0.18	0.08	80.25	89.04	94.62	**89.74**	96.19
UTNetV2 [6]	2022	12.80	15.50	78.97	88.25	94.32	87.60	96.48
FocalUNETR [29]	2023	26.91	16.28	80.37	89.12	94.92	88.66	96.84
LightM-UNet [14]	2024	0.19	0.66	79.24	88.42	94.58	84.95	**97.68**
Ultralight VM-Unet [15]	2024	**0.05**	**0.06**	78.59	88.01	94.30	85.95	96.98
VM-UNet [12]	2024	22.04	4.11	81.35	89.71	94.91	91.12	96.13
U-KAN [30]	2025	9.38	6.89	80.09	88.94	94.60	89.22	96.33
UD-Mamba [31]	2025	19.12	5.91	**81.94**	89.15	94.60	89.55	96.26
MAFUNet [17]	2025	9.61	7.43	81.43	**89.77**	**95.24**	88.89	97.19
Half-MAFUNet	**Proposed Model**	0.53	7.944	80.15	87.68	94.71	83.12	97.99

model complexity, especially when compared with other lightweight architectures that rely primarily on local convolutional features. Several limitations should be noted. The experiments are conducted on three public datasets with fixed train-validation-test splits, which may not fully represent the diversity of clinical data across hospitals, devices, and patient populations. In addition, the main hyperparameter search was performed on CVC-ClinicDB and subsequently applied to BUSI and ISIC-2018, so dataset-specific tuning might further improve performance. Finally, although the model is lightweight, GPU support is still desirable for fast training and inference; very low-power devices may require extra optimization. Despite these limitations, the results imply that combining a compact Half-U-Net backbone with HAM, MAF, CASPP, and an

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

ACA-ASA attention gate is an effective strategy for building accurate yet efficient medical image segmentation models. This design can be adapted to other organs and modalities where both performance and resource constraints are critical. From a practical viewpoint, HALF-MAFUNet's small parameter count and strong accuracy suggest that it can be integrated into computer-aided diagnosis systems running on mid-range GPUs or high-end CPUs in hospitals. Accurate segmentation of colorectal polyps, breast tumors, and skin lesions can support earlier detection, more consistent lesion measurement, and better treatment planning. Future work will build on the current findings by further reducing parameters and FLOPs through more efficient block designs or lightweight attention variants, while preserving segmentation accuracy. In addition, the model will be evaluated on more diverse datasets and multi-center clinical data to better assess its robustness and generalization, and will be integrated into practical computer-aided diagnosis systems for real-world deployment.

V. Conclusion

This study proposed HALF-MAFUNet, a lightweight medical image segmentation model designed to maintain high accuracy while reducing computational cost for practical deployment. The model is built on a Half-U-Net backbone and integrates HAM, MAF, CASPP, and a dual-attention gate comprising ACA and ASA to better fuse global and local features and refine skip connections. Using the optimal hyperparameters (AdamW optimizer, BCE + Dice loss, depth 5, attention ratio 8, and 64 filters), HALF-MAFUNet achieved strong results on three datasets with only about 0.54 million parameters and 7.94 GFLOPs. The model obtained an mIoU/F1- Score of 88.68% / 93.38% on CVC-ClinicDB, 83.83% / 90.50% on BUSI, and 80.15% / 87.68% on ISIC-2018, showing that it can accurately segment colorectal polyps, breast tumors, and skin lesions with low computational complexity. Ablation studies further showed that combining ACA and ASA, with ACA applied first, then ASA, provides the best performance among all attention configurations. Future work will extend this study by evaluating HALF-MAFUNet on larger, multicenter clinical datasets to further assess its robustness and generalization across diverse imaging conditions. Building on the observed efficiency-accuracy trade-off, subsequent research will explore more compact block designs and lightweight attention variants to further reduce parameters and FLOPs, and will also investigates their integration into practical computer-aided diagnosis systems for real-world clinical deployment.

Acknowledgment

We would like to express our sincere gratitude to the Faculty of Information Technology and Data Science, Universitas Sebelas Maret (UNS), Surakarta, Indonesia, for providing High Performance Computing facilities for the implementation of this research. We also extend our deep appreciation to the Intelligence Systems and Humanized Computing Research Group (ISHC-RG UNS), Universitas Sebelas Maret (UNS), for their supervision and guidance in determining the research topic, including mentoring for participation in national competitions, as well as for the execution of this research.

Funding

This work was fully supported and funded by the RKAT Universitas Sebelas Maret (UNS) of the year 2024 under the research grant Hibah Penelitian Fundamental (PF-UNS) with the contract 194.2/UN27.22/PT.01.03/2024.

Author Contribution

Abiaz Fazel Maula Sandy contributed to model implementation, experimental setup, performance evaluation, result analysis, and manuscript drafting. Heri Prasetyo contributed to study supervision, research design, methodological guidance, and critical manuscript revision. All authors read and approved the final manuscript and agreed to be accountable for the accuracy and integrity of the work.

Declarations

Ethical Approval

This study utilizes three publicly available medical image datasets, namely CVC-ClinicDB, BUSI, and ISIC-2018, for segmentation tasks. All datasets consist of anonymized medical images and are openly accessible. Consequently, no additional ethical approval was required for conducting this study.

Consent for Publication Participants.

All participants gave consent for publication

Competing Interests

The authors declare no competing interests.

References

[1] M. G. Linguraru *et al.*, "Clinical, Cultural, Computational, and Regulatory Considerations to Deploy AI in Radiology: Perspectives of RSNA and MICCAI Experts," *Radiol. Artif. Intell.*, vol. 6, no. 4, p. e240225, Jul. 2024, doi: 10.1148/ryai.240225.

[2] J. Zhang *et al.*, "Advances in attention mechanisms for medical image segmentation,"

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Comput. Sci. Rev., vol. 56, p. 100721, May 2025, doi: 10.1016/j.cosrev.2024.100721.

[3] O. Colliot, Ed., *Machine Learning for Brain Disorders*, vol. 197. in Neuromethods, vol. 197. New York, NY: Springer US, 2023. doi: 10.1007/978-1-0716-3195-9.

[4] S. Kolhar and J. Jagtap, "Convolutional neural network based encoder-decoder architectures for semantic segmentation of plants," *Ecol. Inform.*, vol. 64, p. 101373, Sep. 2021, doi: 10.1016/j.ecoinf.2021.101373.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science, vol. 9351. , Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[6] Y. Gao, M. Zhou, D. Liu, Z. Yan, S. Zhang, and D. N. Metaxas, "A Data-scalable Transformer for Medical Image Segmentation: Architecture, Model Efficiency, and Benchmark," 2022, *arXiv*. doi: 10.48550/ARXIV.2203.00131.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," 2017, arXiv. doi: 10.48550/ARXIV.1706.05587.

[8] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, *arXiv*. doi: 10.48550/ARXIV.2010.11929.

[9] R. Wu, Y. Liu, P. Liang, and Q. Chang, "H-vmunet: High-order Vision Mamba UNet for medical image segmentation," *Neurocomputing*, vol. 624, p. 129447, Apr. 2025, doi: 10.1016/j.neucom.2025.129447.

[10] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," 2023, *arXiv*. doi: 10.48550/ARXIV.2312.00752.

[11] J. Jiao *et al.*, "VMamba: Visual State Space Model," in *Advances in Neural Information Processing Systems 37*, Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 103031–103063. doi: 10.52202/079017-3273.

[12] J. Ruan, J. Li, and S. Xiang, "VM-UNet: Vision Mamba UNet for Medical Image Segmentation," 2024, *arXiv*. doi: 10.48550/ARXIV.2402.02491.

[13] J. Xu, "HC-Mamba: Vision MAMBA with Hybrid Convolutional Techniques for Medical Image Segmentation," 2024, *arXiv*. doi: 10.48550/ARXIV.2405.05007.

[14] W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, and L. Ma, "LightM-UNet: Mamba Assists in Lightweight UNet for Medical Image Segmentation," 2024, *arXiv*. doi: 10.48550/ARXIV.2403.05246.

[15] R. Wu, Y. Liu, G. Ning, P. Liang, and Q. Chang, "UltraLight VM-UNet: Parallel Vision Mamba significantly reduces parameters for skin lesion segmentation," *Patterns*, vol. 6, no. 11, p. 101298, Nov. 2025, doi: 10.1016/j.patter.2025.101298.

[16] X. Zhu, W. Wang, C. Zhang, and H. Wang, "Polyp-Mamba: A Hybrid Multi-Frequency Perception Gated Selection Network for polyp segmentation," *Inf. Fusion*, vol. 115, p. 102759, Mar. 2025, doi: 10.1016/j.inffus.2024.102759.

[17] M. Yang, Z. Yang, and N. I. R. Ruhaiyem, "MAFUNet: Mamba with adaptive fusion UNet for medical image segmentation," *Image Vis. Comput.*, vol. 162, p. 105655, Oct. 2025, doi: 10.1016/j.imavis.2025.105655.

[18] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," 2018, *arXiv*. doi: 10.48550/ARXIV.1804.03999.

[19] J. Wang, P. Lv, H. Wang, and C. Shi, "SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in Computed Tomography," *Comput. Methods Programs Biomed.*, vol. 208, p. 106268, Sep. 2021, doi: 10.1016/j.cmpb.2021.106268.

[20] H. Lu, Y. She, J. Tie, and S. Xu, "Half-UNet: A Simplified U-Net Architecture for Medical Image Segmentation," *Front. Neuroinformatics*, vol. 16, p. 911679, Jun. 2022, doi: 10.3389/fninf.2022.911679.

[21] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imaging Graph.*, vol. 43, pp. 99–111, Jul. 2015, doi: 10.1016/j.compmedimag.2015.02.007.

[22] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.

[23] N. Codella *et al.*, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," 2019, *arXiv*. doi: 10.48550/ARXIV.1902.03368.

[24] B. D. Sarira and H. Prasetyo, "Dual Attention and Channel Atrous Spatial Pyramid Pooling Half-UNet for Polyp Segmentation," *J. Electron. Electromed. Eng. Med. Inform.*, vol. 7, no. 3, pp. 680–691, May 2025, doi: 10.35882/jeeemi.v7i3.893.

[25] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark,"

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357

Neurocomputing, vol. 503, pp. 92–108, Sep. 2022, doi: 10.1016/j.neucom.2022.06.111.
[26] M. Lee, "Mathematical Analysis and Performance Evaluation of the GELU Activation Function in Deep Learning," *J. Math.*, vol. 2023, pp. 1–13, Aug. 2023, doi: 10.1155/2023/4229924.
[27] J. M. J. Valanarasu and V. M. Patel, "UNeXt: MLP-Based Rapid Medical Image Segmentation Network," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, vol. 13435, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., in Lecture Notes in Computer Science, vol. 13435. , Cham: Springer Nature Switzerland, 2022, pp. 23–33. doi: 10.1007/978-3-031-16443-9_3.
[28] J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu, "MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA: IEEE, Dec. 2022, pp. 1150–1156. doi: 10.1109/BIBM55620.2022.9995040.
[29] C. Li *et al.*, "FocalUNETR: A Focal Transformer for Boundary-aware Segmentation of CT Images," 2022, *arXiv*. doi: 10.48550/ARXIV.2210.03189.
[30] C. Li *et al.*, "U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation," 2024, *arXiv*. doi: 10.48550/ARXIV.2406.02918.
[31] W. Zhao, F. Wang, Y. Wang, Y. Xie, Q. Wu, and Y. Zhou, "UD-Mamba: A pixel-level uncertainty-driven Mamba model for medical image segmentation," 2025, *arXiv*. doi: 10.48550/ARXIV.2502.02024.

Author Biography

Abiaz Fazel Maula Sandy is a final-year undergraduate student at Universitas Sebelas Maret (UNS), majoring in Informatics under the Faculty of Information Technology and Data Science. With a strong passion for data technologies and artificial intelligence, he has pursued his academic journey closely tied to both research and practical implementation in computer science. He won the national competition for scientific writing on IT-related topics at GEMASTIK 2025. He is particularly interested in developing machine learning and deep learning-based solutions for real-world challenges. In addition to his studies, he actively seeks opportunities to collaborate, grow professionally, and contribute to high-impact research. He can be contacted at abiazfazel_ms@student.uns.ac.id.

Heri Prasetyo received the doctoral degree from the Department of Electrical Engineering, National Taiwan University of Science and Technology (NTUST), Taiwan, in 2015. He received the Best Dissertation Award from the Taiwan Association for Consumer Electronics (TACE) in 2015, the Best Paper Awards from the International Symposium on Electronics and Smart Devices 2017 (ISESD 2017), ISESD 2019, the International Conference on Science in Information Technology (ICSITech, 2019), the International Conference on Smart Technology, Applied Informatics, and Engineering (APICS 2022), the International Conference on Informatics and Computing (ICIC 2023), International Conference on Computer, Control, Informatics and its Applications (IC3INA 2024), International Conference on Electronics Representation and Algorithm (ICERA 2025), International Conference on Artificial Intelligence Future Implementations (ICAIFI 2025), and the Outstanding Faculty Award from Universitas Sebelas Maret (UNS) in 2019 and 2023. His research interests include multimedia signal processing, computational intelligence, pattern recognition, and machine learning. He can be contacted at heri.prasetyo@staff.uns.ac.id.

Manuscript received 12 September 2025; Revised 15 November 2025; Accepted 5 January 2026; Available online 12 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1357