Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 1, January 2026, pp: 206-221 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Medical Image Segmentation Using a Global Context-Aware and Progressive Channel-Split Fusion U-Net with Integrated Attention Mechanisms

Alfath Roziq Widhayaka, and Heri Prasetyo

Department of Informatics, Faculty of Information Technology and Data Science, Universitas Sebelas Maret (UNS), Surakarta, Indonesia

Corresponding author: Heri Prasetyo (e-mail: heri.prasetyo@staff.uns.ac.id), **Author(s) Email**: Alfath Roziq Widhayaka (alfathroziq94@student.uns.ac.id)

Abstract **Medical image segmentation serves as a key component in Computer-Aided Diagnosis (CAD) systems across various imaging modalities. However, the task remains challenging because many images have low contrast and high lesion variability, and many clinical environments require efficient models. This study proposes CFCSE-Net, a U-Net-based model that builds upon X-UNet as a baseline for the CFGC and CSPF modules. This model incorporates a modified CFGC module with added Ghost Modules in the encoder, a CSPF module in the decoder, and Enhanced Parallel Attention (EPA) in the skip connections. The main contribution of this paper is the design of a lightweight architecture that combines multi-scale feature extraction with an attention mechanism to maintain low model complexity and increase segmentation accuracy. We train and evaluate CFCSE-Net on four public datasets: Kvasir-SEG, CVC-ClinicDB, BUSI (resized to 256 × 256 pixels), and PH2 (resized to 320 × 320 pixels), with data augmentation applied. We report segmentation performance as the mean ± standard deviation of IoU, DSC, and accuracy across three random seeds. CFCSE-Net achieves 79.78% ± 1.99 IoU, 87.21% ± 1.72 DSC, and 96.70% ± 0.59 accuracy on Kvasir-SEG, 88.11% ± 0.86 IoU, 93.42% ± 0.55 DSC, and 99.04% ± 0.09 accuracy on CVC-ClinicDB, 69.33% ± 2.66 IoU, 78.80% ± 2.65 DSC, and 96.30% ± 0.51 accuracy on BUSI, and 92.27% ± 0.52 IoU, 95.92% ± 0.30 DSC, and 98.06% ± 0.16 accuracy on PH2. Despite its strong performance, the model remains compact with 909,901 parameters and low computational cost, requiring 3.24 GFLOPs for 256 × 256 inputs and 5.07 GFLOPs for 320 × 320 inputs. These results show that CFCSE-Net maintains stable performance on polyp, breast ultrasound, and skin lesion segmentation while it stays compact enough for CAD systems on hardware with low computational resources.**

Keywords **Deep Learning; Medical image segmentation; Efficient neural architectures; CFCSE-Net**

I. Introduction

Medical image segmentation holds a crucial role in research, especially in the development of computer-aided diagnosis (CAD) systems [1]. Advances in imaging technology enable researchers to create various methods that support clinical diagnosis, biomedical studies, and visual information analysis [2], [3]. In general, medical image segmentation separates anatomical structures and pathological regions across various image modalities, enabling physicians to analyze lesions more accurately and obtain reliable diagnostic information [4]. The manual process of medical image segmentation faces several challenges, such as low image contrast that makes object boundaries hard to identify, organ and lesion shapes vary, and differences in image quality caused by device and configuration variations [5]. These conditions increase the risk of errors in manual segmentation because medical personnel experience fatigue, subjective interpretation, and inconsistent skill levels [6]. Researchers, therefore, adopt deep learning, especially Convolutional Neural Networks (CNNs), to mitigate these problems. CNNs are able to capture complex image patterns automatically, make manual feature design less important, and produce better segmentation accuracy [7].

The U-Net architecture by [8] plays a central role in medical image segmentation. U-Net uses a U-shaped encoder–decoder structure that learns context and restores spatial detail even when the dataset has few annotations. The U-Net architecture has been extended to produce several variants, including U-Net++, introduced in [9], which improves the connection path and has been shown to increase accuracy across various medical image modalities. However, this improvement requires high

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

computational resources. To address computational efficiency, [10] proposed the Half-UNet with an asymmetric design that reduces redundancy in the decoder path and channel usage. This method combines multi-scale features from UNet3+ and uses ghost modules to increase efficiency while keeping a uniform number of channels at each level. Ghost modules produce additional feature maps through simple operations on the main feature map, which improves efficiency [11]. With this design, Half-UNet reduces the number of parameters by up to 98.6% while maintaining segmentation performance comparable to that of U-Net and its variants.

Several recent works aim to improve medical image segmentation accuracy by using attention mechanisms that highlight important information while keeping the model efficient. The study in [12] proposes BRAU-Net++, which combines the U-Net structure with Bi-Former (Transformer Attention) and Selective Cross-Channel Spatial Attention (SCCSA). This combination allows BRAU-Net++ to balance global and local information and reduce background interference. Furthermore, [13] introduces One-Point-Five U-Net, a segmentation framework based on the U-Net architecture that employs Enhanced Parallel Attention (EPA) to enable the network to capture global and local information in parallel under noisy conditions. Overall, attention mechanisms help the network remove complex noise and make it easier to detect weak or low-contrast tissue edges, so the segmentation results stay stable.

Recent studies show that multi-scale processing increases segmentation accuracy by helping the model combine global context with fine local information. In this context, X-UNet [14] introduces a collaborative fusion framework that serves as the baseline architecture of this work. X-UNet integrates Collaborative Fusion with Global Context-Aware (CFGC) modules to extract multi-scale contextual information and Cross Split-Channel Progressive Fusion (CSPF) modules to align encoder and decoder features through channel splitting and progressive fusion. By explicitly coordinating global context extraction and progressive feature alignment, X-UNet improves information flow across network stages and enhances segmentation performance on diverse medical imaging datasets. However, the original X-UNet relies on standard convolutions and uniform feature fusion, which may introduce parameter redundancy and limit efficiency when deployed in resource-constrained settings. In contrast to X-UNet, other context-aware approaches, such as CANet [15], introduce a context-aware network that uses two pyramidal pipelines with L-shaped designs to generate multi-resolution inputs and multi-scale convolutions with chained residual pooling. These components highlight important features, reduce noise, and improve feature fusion between the encoder and decoder. They also emphasize the importance of global context modeling in medical image segmentation.

Based on findings from prior studies, we propose an improved model architecture, Collaborative Fusion with Global Context-Aware and Cross-Split-Channel Progressive Fusion with Enhanced Parallel Attention Network (CFCSE-Net). The model aims to improve accuracy by incorporating multi-scale features and an attention mechanism, while maintaining computational efficiency to keep the model lightweight. Building upon the CFGC and CSPF modules of X-UNet, CFCSE-Net incorporates three main modifications, including (1) the encoder replaces regular convolution with a modified CFGC module that includes a Ghost Module; (2) the skip connections use Enhanced Parallel Attention (EPA) to strengthen feature transfer from the encoder to the CSPF; and (3) the decoder replaces regular convolution with the CSPF module only to reduce architectural complexity and focus the decoder on progressive feature fusion, while preserving segmentation accuracy. This design is expected to be lightweight, with a small parameter size and low GFLOPs consumption, while still providing high accuracy in medical image segmentation, thereby supporting disease prevention and diagnosis more effectively.

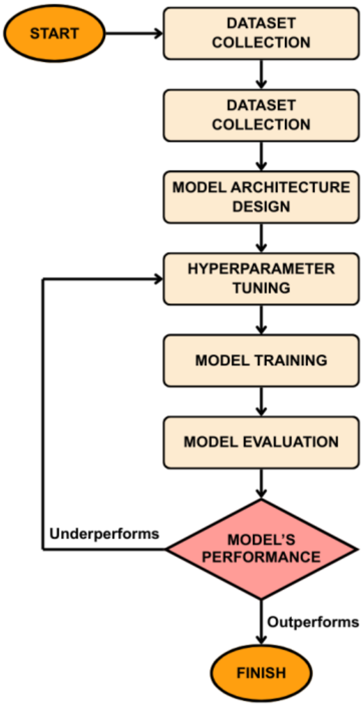


Fig. 1. Stages of the proposed research process.

II. Methodology

Fig. 1 presents the overall workflow of this study. The research process follows six main stages: (1) dataset

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i1.1371

collection, (2) dataset preprocessing, (3) model architecture design, (4) hyperparameter tuning, (5) model training, and (6) model evaluation.

A. Dataset Collection

This study uses four open-source medical image datasets for the segmentation task. The Kvasir-SEG offers 1,000 gastrointestinal endoscopy images with different masks and resolutions [16]. The CVC-ClinicDB has 612 colonoscopy images with a native size of 288 × 368 pixels [17]. The Breast Ultrasound Images (BUSI) dataset contains 780 breast ultrasound images [18]; and this study uses 647 benign and malignant images because the normal class does not contain an ROI. The PH^2 dataset provides 200 dermatoscopic images of skin lesions along with their annotations [19].

B. Dataset Preprocessing

After data collection, this study splits each dataset into training and testing sets with an 80:20 ratio. The study repeats the splitting process using three different random seeds to reduce split bias and ensure that the reported performance does not depend on a single favorable data partition, thereby improving reproducibility and result stability across different data splits. The study also standardizes the image sizes to match the architectural input. This normalization process prepares all datasets for stable training and prevents scale differences from affecting feature extraction. Images from Kvasir-SEG, CVC-ClinicDB, and BUSI are resized to 256 × 256 pixels, whereas PH2 images are resized to 320 × 320 pixels due to their higher native resolution and finer dermoscopic lesion detail. For the other datasets, 256 × 256 offers a good balance between spatial detail and computational efficiency. Thus, input resolution is chosen based on dataset-specific spatial characteristics rather than treated as an independent experimental variable.

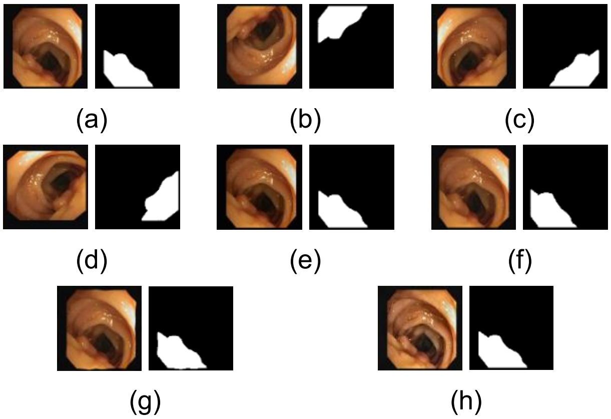


Fig. 2. Examples of data augmentation results: (a) original, (b) vertical flip, (c) horizontal flip, (d) random rotation, (e) random brightness, (f) grid distortion, (g) elastic deformation, and (h) CLAHE contrast enhancement.

In the training set, the study applies data augmentation eight times to increase sample variation and reduce overfitting [20]. The augmentations include horizontal and vertical flips, random rotations (90 °, 180 °, 270 °), random brightness/contrast adjustments, grid distortion, elastic deformation, and CLAHE contrast enhancement. Specifically, each original image–mask pair is transformed into eight variants (including the original), each generated by applying a single augmentation operation. This augmentation strategy expands the visual variability of the dataset, improves the stability of feature learning, and supports better generalization during segmentation [21]. Fig. 2 provides examples of the augmentation results, and Table 1 summarizes the dataset distribution.

Table 1. Split dataset distribution.

Dataset	Training Set	Training Set (Augmented)	Test Set
Kvasir-SEG	800	6400	200
CVC-ClinicDB	489	3912	123
BUSI	517	4136	130
PH^2	160	1280	40

C. Model Architecture Design

The proposed CFCSE-Net architecture is illustrated in Fig. 3 and represents an extended version of the standard U-Net framework through several structural refinements. The network uses an encoder–decoder structure and is based on the X-UNet architecture [14], which integrates modified CFGC and CSPF modules. CFCSE-Net applies a channel configuration of (16, 32, 64, 128, 256). On the encoder, the model uses the CFGC module to extract multi-scale features and apply global context weighting. The model also incorporates Ghost Modules into the CFGC module to keep informative feature representations while reducing computational cost [11]. From each encoder block, the model sends skip connections to the decoder before the downsampling step. For every skip connection, the model applies the Enhanced Parallel Attention (EPA) mechanism to emphasize important channel and spatial information while reducing noise. The model then sends the refined skip features to the CSPF module on the decoder path. In the decoder, CSPF aligns the incoming features and fuses them channel-wise with its internal representations. This process preserves boundary details and global contextual cues and improves segmentation accuracy. The conventional convolutional blocks in the encoder are replaced with a modified Collaborative Fusion with Global Context-Aware

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i1.1371

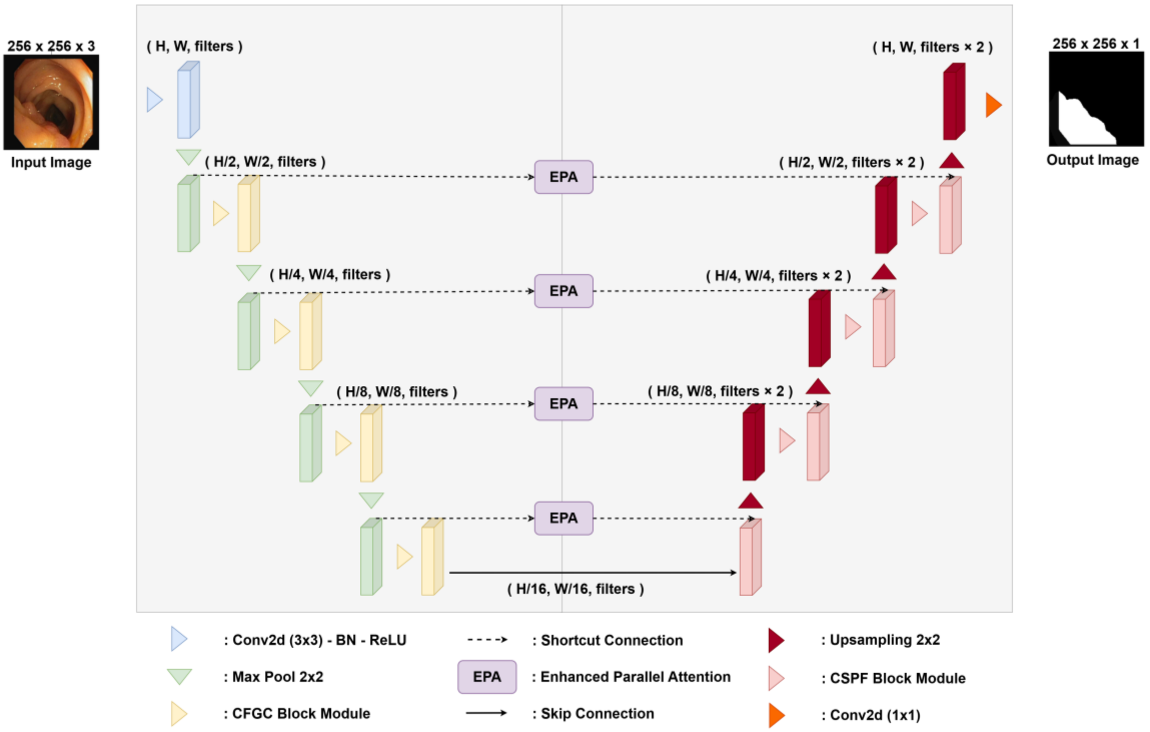


Fig. 3. Architecture of the proposed CFCSE-Net.

(CFGC) module with Ghost Modules, as shown in Fig. 4 (a). Compared to the original CFGC module in X-UNet [14], the proposed CFGC in CFCSE-Net preserves the original fusion and global context modelling scheme. The model modifies only the convolutional operations to improve parameter efficiency. The proposed CFGC module replaces the original 3×3 and 7×7 depthwise convolutions with 3×3 and 7×7 Ghost depthwise convolutions replaces the 1×1 standard convolution with a 1×1 Ghost convolution, and replaces the final 3×3 standard convolution with a 3×3 Ghost convolution. Both GhostDWConv and GhostConv follow the Ghost Module principle, generating additional feature maps from a small set of intrinsic features using inexpensive linear operations. GhostDWConv is applied during primary feature extraction to efficiently capture multi-scale spatial patterns, whereas GhostConv performs channel regulation and feature refinement after multi-scale fusion. By integrating these Ghost Modules, the CFGC module achieves significant reductions in parameter count and computational cost without altering its original fusion structure or global context modeling capability.

The process begins with an input feature map $X_1 \in R^{C \times H \times W}$, where $C, H,$ and W denote the channel, height, and width dimensions. The CFGC module applies Ghost depthwise convolutions with kernel sizes $k \in \{3,7\}$ to extract multi-scale feature representations and produce two output feature maps X_2 and X_3, which correspond to local and global receptive-field scales, respectively, as illustrated in Fig. 4 (a). The 3×3 GhostDWConv captures fine local spatial details with a limited receptive field, while the 7×7 GhostDWConv captures long-range contextual and structural information. This operation is defined in Eq. (1) [14].

$$X_i = GhostDWConv_{k \times k}(X_1), \qquad k \in \{3,7\} \tag{1}$$

The module then concatenates these multi-scale features and applies a 1×1 Ghost convolution to regulate the channel size and computational cost, generating the fused feature map X_4, as expressed in Eq. (2) [14].

$$X_4 = GhostConv_{1 \times 1}(Concat(X_2, X_3)) \tag{2}$$

To capture structural patterns distributed along the vertical (H) and horizontal (W) dimensions, CFGC computes channel-wise average and max projections from X_2 and X_3. These are then combined into pairs of directional descriptors H_1 and W_1, which are further fused using 3×3 convolutions and reshaped, as shown in Eqs. (3) and (4), following [14], where H and W represent the aggregated vertical and horizontal attention features.

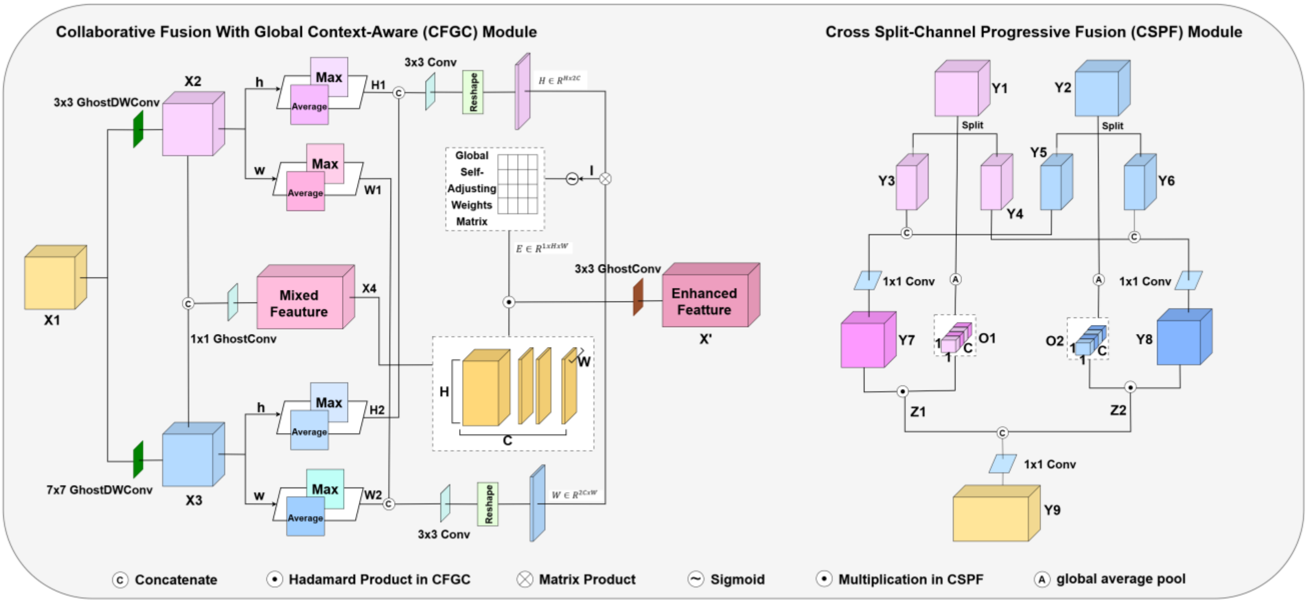
$$H = Reshape(Conv_{3 \times 3}(Concat(H_1, H_2))) \tag{3}$$

$$W = Reshape(Conv_{3 \times 3}(Concat(W_1, W_2))) \tag{4}$$

By performing matrix interactions between (H_1, H_2) and (W_1, W_2), the module approximates global spatial self-attention and computes a set of global self-adjusting weights $I \in R^{1 \times H \times W}$, as defined in Eq. (5) [14].

$$I_{i,j} = \frac{\exp\,(H_i * W_j)}{\sum_{i=1}^{n} \exp\,(H_i * W_j)'}, \qquad n = 2C \times 2C \tag{5}$$

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371



(a) (b)

Fig. 4. Architecture of (a) Collaborative Fusion with Global Context-Aware (CFGC) module and (b) Cross split-channel progressive fusion (CSPF) module.

The module applies a sigmoid function to normalize the weights and generates a global response map E, as shown in Eq. (6) [14], where E represents the spatial attention response.

$$E = sigmoid(I) \tag{6}$$

The response map then reweights X_4 through element-wise calibration. The final output X' is obtained via 3×3 Ghost convolution, as expressed in Eq. (7) [14].

$$X' = GhostConv_{3\times3}(X_4 * E) \tag{7}$$

Through this design, CFGC effectively combines local details (from 3×3 Ghost depthwise filters) and global contextual cues (from 7×7 Ghost depthwise filters and spatial weighting) while maintaining a low parameter footprint. This enables the model to remain lightweight yet capable of capturing long-range dependencies that are crucial for segmenting ambiguous anatomical boundaries and wide structural regions. On the decoder side, the model replaces standard convolutional blocks with the Cross Split-Channel Progressive Fusion (CSPF) module, as illustrated in Fig. 4 (b). CSPF progressively aligns and fuses encoder–decoder features in the channel domain to reduce semantic discrepancies and strengthen feature consistency. The first step involves feeding the encoder feature Y_1 and decoder feature Y_2. Both feature maps are evenly divided along the channel dimension, resulting in four sub-features Y_3, Y_4, Y_5, and Y_6, where Y_3 and Y_4 denote the split encoder features, while Y_5 and Y_6 denote the split decoder features. The cross-paired groups (Y_3, Y_5) and (Y_4, Y_6), are concatenated and fused using 1×1 convolutions to capture cross-level correspondence, as described in Eqs. (8) and (9), following [14].

$$Y_7 = Conv_{1\times1}(Concat(Y_3, Y_5)) \tag{8}$$

$$Y_8 = Conv_{1\times1}(Concat(Y_4, Y_6)) \tag{9}$$

To make the fusion channel-aware, channel weighting vectors are generated via global average pooling on Y_1 and Y_2, as defined in Eq. (10) [14], where O_1 and O_2 are derived from the encoder and decoder features, respectively.

$$O_i = AveragePool(Y_i), \qquad i \in \{1,2\} \tag{10}$$

These weighting vectors modulate Y_7 and Y_8 through element-wise multiplication, producing calibrated feature sets Z_1 and Z_2, as expressed in Eq. (11) [14], where Z_1 and Z_2 represent the channel-attended fused features.

$$Z_1 = Y_7 * O_1, \qquad Z_2 = Y_8 * O_2 \tag{11}$$

Finally, the calibrated features are concatenated and compressed using a 1×1 convolution to yield the fused output Z', as described in Eq. (12) [14].

$$Z' = Conv_{1\times1}(Concat(Z_1, Z_2)) \tag{12}$$

With this mechanism, CSPF is able to align detailed information from the encoder and context from the decoder more adaptively on each channel. To further reduce computational cost, the architecture incorporates the Ghost Module within the CFGC block, with the structure of the Ghost Module illustrated in Fig. 5. This module is applied at several stages to enhance feature extraction efficiency by lowering both parameter count and computational load. Inspired by GhostNet [11], it replaces standard convolution with two lightweight operations: a pointwise convolution that generates the primary features and a depthwise convolution that

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

produces additional "ghost features." This approach enables the network to obtain rich feature representations while maintaining significantly lower computational complexity compared with conventional convolutional layers.

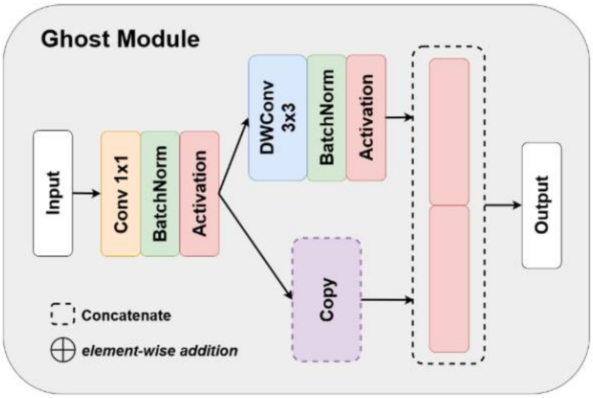


Fig. 5. Architecture of Ghost Module.

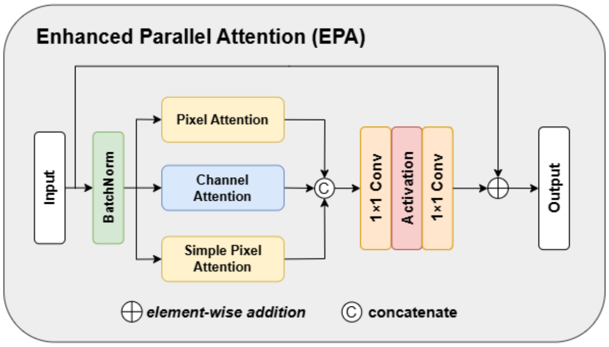


Fig. 6. Structure of the Enhanced Parallel Attention (EPA).

The network places the Enhanced Parallel Attention (EPA) module in the skip-connection path to strengthen important features before they enter the decoder stage. The structure of the EPA module is illustrated in Fig. 6. EPA has three parallel attention branches: Simple Pixel Attention (SPA), Channel Attention (CA), and Pixel Attention (PA). CA captures global channel relationships, whereas SPA and PA emphasize spatial information at fine spatial scales [22]. The outputs of the three branches are stacked along the channel axis, and the network sends them to a PWConv–ReLU–PWConv block to obtain the required channel size. The model also adds an identity shortcut to preserve original information and maintain stable gradient flow. Through this design, EPA highlights relevant features, reduces noise, and improves segmentation accuracy.

D. Hyperparameter Tuning

In this study, hyperparameter tuning is performed on the CVC-ClinicDB dataset, which exhibits representative variability in lesion size, shape, and boundary clarity across the evaluated datasets. This dataset presents sufficient complexity to reflect common segmentation challenges while maintaining stable training behavior, making it suitable for identifying robust hyperparameter configurations. The tuning process aims to determine the most effective parameter combination that improves model performance on the segmentation task [23]. It evaluates key training components, including activation functions, optimizers, and loss functions, using a controlled empirical evaluation approach through comparative experiments to observe their impact on training stability and segmentation quality. During hyperparameter tuning, one hyperparameter category is varied at a time while others remain fixed.

The study tests several activation functions such as ReLU, LeakyReLU, ELU, SELU, and GELU, to analyze their influence on the learning process. The study also tests various optimizers, including Adam, NAdam, AdamW, and RMSProp, to observe their convergence behavior during training. In addition, the study evaluates several loss functions, such as Binary Cross-Entropy (BCE), Dice Loss, Tversky Loss, and the combined BCE + Dice Loss, to determine the loss formulation that produces stable and accurate segmentation results. The selection of optimal hyperparameters relies on quantitative performance metrics, where configurations that produce higher Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and Accuracy (Acc) values indicate optimal performance.

E. Model Training

The model training phase uses the optimal hyperparameters obtained from the tuning process. The training procedure takes the training data as input and produces segmentation masks as output. This work uses the CVC-ClinicDB dataset for hyperparameter tuning and ablation studies because it has reliable annotations and a moderate level of difficulty. The training phase uses RGB medical images as input and binary masks as output for all datasets. A high-performance workstation with an NVIDIA RTX A4000 GPU runs all experiments, and the PyTorch framework handles model implementation and optimization.

F. Model Evaluation

For evaluation, this study uses the test set of each dataset to measure segmentation quality. The analysis reports three main metrics: Intersection over Union (IoU), Dice Similarity Coefficient (DSC), and accuracy, which compare model predictions with the ground truth [24]. The study also checks computational efficiency using the number of parameters and GFLOPs to describe model size and computational cost. The Intersection over Union (IoU) metric in Eq. (13) measures the overlap between the region the model predicts as positive and the ground-truth region [25].

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i1.1371

Table 2. Comparison of the initial hyperparameters with the resulting optimal configuration.

Hyperparameter	Initial	Optimal
Batch Size	4	4
Seed	3	3
Epoch	100	100
Image Size	256×256 (Kvasir-SEG, CVC-ClinicDB, & BUSI)	256×256 (Kvasir-SEG, CVC-ClinicDB, & BUSI)
	320×320 (PH^2)	320×320 (PH^2)
Learning Rate	0.0003 (Kvasir-SEG, CVC-ClinicDB, & PH^2)	0.0003 (Kvasir-SEG, CVC-ClinicDB, & PH^2)
	0.0005 (BUSI)	0.0005 (BUSI)
Cosine Annealing Delay	0.00001	0.00001
Optimizer	Adam	NAdam
Activation Function	ReLU	ReLU
Loss Function	BCE + DL	BCE + DL

$$IoU = \frac{TP}{TP + FP + FN} \tag{13}$$

In this equation, TP denotes true positive pixels, FP denotes pixels that the model predicts as positive but belong to the negative class, and FN denotes positive pixels that the model does not detect. The Dice coefficient in Eq. (14) quantifies the overlap between the predicted and ground-truth masks [25].

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{14}$$

This metric emphasizes the balance between correctly detected positive pixels and errors produced by FP and FN. The accuracy metric in Eq. (15) calculates the proportion of correctly predicted pixels [25].

$$Acc = \frac{TN + TP}{TN + TP + FN + FP} \tag{15}$$

This equation uses TN to denote true-negative pixels, i.e., pixels in the negative class that the model predicts as negative. This study calculates the number of trainable parameters using Eq. (16) [26].

$$Params = K_h \times K_w \times C_{in} \times C_{out} \tag{16}$$

Here, K denotes the kernel size, C_{in} represents the number of input channels, and C_{out} refers to the number of output channels. GFLOPs are defined in Eq. (17) to measure the computational workload of the model [26].

$$GFLOPs = \frac{2 \times (K^2 \times C_{in} \times C_{out} \times H_{out} \times W_{out})}{10^9} \tag{17}$$

Here, H_{out} and W_{out} denote the output height and width. GFLOPs indicate the number of operations for one forward pass.

III. Result

A. Hyperparameter Tuning

This study performs hyperparameter tuning by testing several configuration choices, including different optimizers, activation functions, and loss functions. The CVC-ClinicDB dataset serves as the basis for tuning and identifies the most effective settings for model development. All experiments were run sequentially, allowing each run to refine the previous configuration and progressively converge towards more optimal settings. Table 2 presents a comparison between the initial hyperparameters with the optimized values derived incrementally throughout the overall successive stages of the tuning process. The initial experiments determine the appropriate optimizer to balance model performance and computational complexity. Table 3 presents the results of this stage, and Fig. 7 compares the last 5 epochs. The study evaluates four optimizers on the CVC-ClinicDB dataset: Adam, NAdam, AdamW, and RMSProp. NAdam achieves the highest consistency in accuracy and is selected as the primary optimizer for subsequent experiments. The second experiment evaluates five activation functions, namely ReLU, LeakyReLU, ELU, SELU, and GELU, while keeping NAdam as the fixed optimizer. Table 3 shows that the choice of activation function significantly affects accuracy, although it does not change the total number of model parameters. ReLU provides the highest performance, so this study uses it as the default activation function for later experiments. The last experiment compares four loss functions under imbalanced-class conditions in medical image segmentation. These loss functions include Binary Cross-Entropy (BCE), Dice Loss (DL), Tversky Loss (TL), and a combined Binary Cross-Entropy and Dice

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i1.1371

Loss (BCE + DL). Table 3 shows that BCE + DL achieves the highest accuracy; therefore, this study adopts it as the loss function for training the proposed model, as it integrates pixel-level supervision with region-overlap optimization to enable stable and discriminative feature learning.

B. Model Ablation

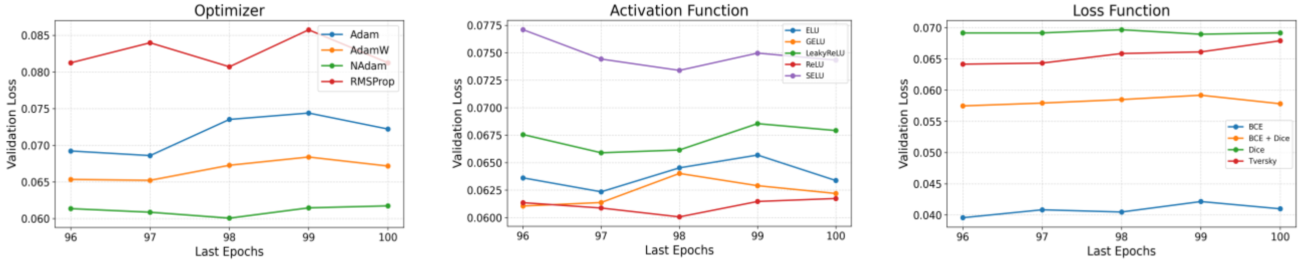
The ablation experiment examines how each component of the CFCSE-Net architecture contributes to performance by selectively removing or modifying specific modules. The model integrates three main modules: the CFGC Module, the CSPF Module, and the EPA Module. The study evaluates seven configurations: (1) U-Net baseline, (2) CFCSE-Net without CSPF and EPA, (3) CFCSE-Net without CFGC and EPA, (4) CFCSE-Net without CSPF, (5) CFCSE-Net without CFGC, (6) CFCSE-Net without EPA, and (7) the full CFCSE-Net as the proposed method.

Table 4 shows that each module provides a distinct contribution to segmentation performance, while Table 5 provides a qualitative comparison of the ablation variants by visually contrasting the predicted masks with the input images and ground truth. Removing the CSPF module (Ablation 4) reduces IoU and DSC, which indicates that CSPF improves feature representation through progressive channel-split fusion in the decoder. This fusion mechanism aligns and combines multi-scale features more effectively, allowing the decoder to recover fine-grained spatial details. Removing the CFGC module (Ablation 5) results in a larger performance drop, confirming that global context information plays a critical role in guiding encoder feature extraction. By modeling long-range dependencies, CFGC helps the network distinguish target regions from complex backgrounds and improves boundary consistency.

Removing the EPA module (Ablation 6) results in a moderate performance decrease, indicating that EPA enhances information flow across skip connections by emphasizing relevant encoder features prior to fusion. Configurations that remove two modules simultaneously (Ablation 2 and Ablation 3) show more severe performance degradation, which demonstrates that the modules complement each other rather than operate

Table 3. Hyperparameter tuning results.

Hyperparameter		IoU/%	DSC/%	Acc/%	Parameter	GFLOPs
Optimizer	Adam	86.07	91.73	98.91	909,901	3.24G
	NAdam	**87.64**	**93.17**	**99.01**	**909,901**	**3.24G**
	AdamW	85.76	91.54	98.90	909,901	3.24G
	RMSProp	83.94	90.59	98.51	909,901	3.24G
Activation Function	**ReLU**	**88.04**	**93.31**	**99.02**	**909,901**	**3.24G**
	LeakyReLU	87.30	92.56	98.96	909,901	3.24G
	ELU	86.62	92.01	98.98	909,901	3.24G
	SELU	85.36	91.63	98.80	909,901	3.24G
	GELU	87.10	92.68	99.00	909,901	3.24G
Loss Function	BCE	86.28	91.93	98.90	909,901	3.24G
	DL	86.95	92.12	99.02	909,901	3.24G
	Tversky Loss	85.92	91.60	98.89	909,901	3.24G
	BCE + DL	**88.04**	**93.31**	**99.02**	**909,901**	**3.24G**



(a) (b) (c)

Fig. 7. The hyperparameter tuning results in the last 5 epochs are presented in graphs of: (a) optimizer, (b) activation function, and (c) loss function.

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

independently. Overall, the full CFCSE-Net model achieves the best results with an IoU of 88.04%, a DSC of 93.31%, an accuracy of 99.02%, 0.90M parameters, and a complexity of 3.24 GFLOPs. These results show that the combination of CFGC, CSPF, and EPA offers the best balance between segmentation accuracy and computational cost.

C. Experiment Result

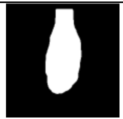
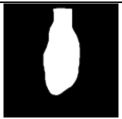
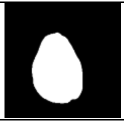
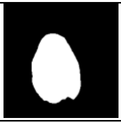
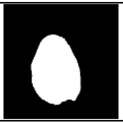
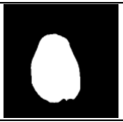
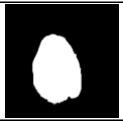
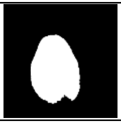
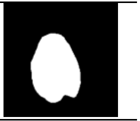
After the hyperparameter tuning stage and the ablation study, this work selects the optimal parameter settings and the most effective model variant. To improve reproducibility and reduce potential bias from a single data split, each dataset is evaluated using three random train-test splits, and the final results are reported as the mean and standard deviation of the performance metrics. Table 2 lists the optimal hyperparameters, which are consistently applied during training on the Kvasir-SEG, CVC-ClinicDB, BUSI, and PH2 datasets. Table 6 summarizes the quantitative performance of CFCSE-Net across all evaluated datasets. The model achieves an IoU of 79.78% ± 1.99 on Kvasir-SEG, 88.11% ± 0.86 on CVC-ClinicDB, 69.33% ± 2.66 on BUSI, and 92.27% ± 0.52 on PH2. Correspondingly, the DSC values reach 87.21% ± 1.72, 93.42% ± 0.55, 78.80% ± 2.65, and 95.92% ± 0.30, while the accuracy remains above 96% across all evaluated datasets. The relatively small standard deviations indicate that the proposed model exhibits stable and robust performance across different random splits. In terms of efficiency, CFCSE-Net maintains a lightweight design with approximately 0.90 million parameters and low computational complexity, requiring 3.24 GFLOPs for most datasets and 5.07 GFLOPs for PH2, which has a higher input resolution. These results demonstrate that the proposed model achieves strong segmentation accuracy while preserving computational efficiency across different imaging modalities.

During training, this study tracks both training loss and validation loss to analyze convergence behavior and training stability, as shown in Fig. 8. The figure illustrates representative loss curves from one of the three random train–test splits for each dataset. For all datasets, the training loss consistently decreases and converges smoothly, indicating stable optimization. The validation loss follows a similar trend and remains close to the training loss, suggesting that the model does not suffer from severe overfitting. Slight fluctuations in the validation loss, particularly BUSI, reflect the higher variability and complexity of ultrasound images rather than unstable training. Overall, the loss curves confirm that CFCSE-Net converges effectively and generalizes well across different datasets. After training, the study evaluates the trained model on each dataset's test set to assess generalization performance. Table 6 reports the quantitative test results, while Table 7 presents qualitative visual comparisons between the predicted segmentation masks and the ground-truth masks. CFCSE-Net shows close alignment with the ground truth on Kvasir-SEG, CVC-ClinicDB, and PH^2, where clearer boundaries, higher contrast, and more homogeneous

Table 4. **Ablation Model Results.**

Method	IoU/%	DSC/%	Acc/%	Parameter	GFLOPs
Ablation 1	87.48	92.70	99.00	1,944,049	6.93G
Ablation 2	86.49	92.26	98.89	1,122,225	4.50G
Ablation 3	87.66	92.86	99.00	1,444,289	3.87G
Ablation 4	85.95	91.46	98.88	1,213,101	5.08G
Ablation 5	84.15	90.60	98.66	**748,013**	2.92G
Ablation 6	87.77	93.19	99.00	819,025	**2.66G**
CFCSE-Net	**88.04**	**93.31**	**99.02**	909,901	3.24G

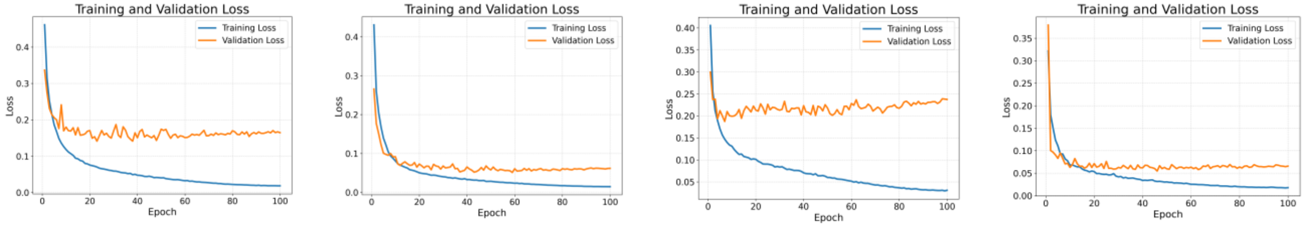
Table 5. **Qualitative comparison of model ablation (a) input image, (b) ablation 1, (c) ablation 2, (d) ablation 3, (e) ablation 4, (f) ablation 5, (g) ablation 6, (h) CFCSE-net, (i) ground truth**

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
								
								

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

Table 6. Performance results obtained from all four datasets.

Dataset	IoU/%	DSC/%	Acc/%	Parameter	GFLOPs
Kvasir-SEG	79.78 ± 1.99	87.21 ± 1.72	96.70 ± 0.59	909.901	3.24G
CVC-ClinicDB	88.11 ± 0.86	93.42 ± 0.55	99.04 ± 0.09		
BUSI	69.33 ± 2.66	78.80 ± 2.65	96.30 ± 0.51		
PH^2	92.27 ± 0.52	95.92 ± 0.30	98.06 ± 0.16		5.07G



(a) (b) (c) (d)

Fig. 8. Training and validation loss curves for (a) Kvasir-SEG, (b) CVC-ClinicDB, (c) BUSI, and (d) PH^2.

backgrounds allow the CFGC, CSPF, and EPA modules to effectively capture global context and preserve boundary details, resulting in smooth and accurate segmentation. In contrast, on the BUSI dataset, although the model successfully identifies the overall lesion regions, minor boundary inaccuracies arise from ultrasound-specific challenges, such as speckle noise, low contrast, and heterogeneous tissue structures. These qualitative observations are consistent with the quantitative results and indicate that ultrasound-adaptive feature enhancement could further improve segmentation performance in future work.

IV. Discussion

This study introduces CFCSE-Net, a segmentation architecture that improves accuracy while maintaining computational efficiency. The model evaluates four heterogeneous datasets and compares the results with several well-established architectures, including U-Net [8], U-Net++ [9], AttU-Net [27], TransUNet [28], UCTransUNet [29], CANet [15], TransAttUNet [30], BRAUNet++ [12], FDFUNet [31], and X-UNet [14], which serves as the primary architectural baseline for the CFGC and CSPF modules. Compared to X-UNet, CFCSE-Net introduces targeted modifications to enhance efficiency and feature utilization. In the encoder, the model replaces standard convolutions in the CFGC module with Ghost-based operations to reduce parameter redundancy while preserving global context modelling. Along the skip connections, the model integrates the Enhanced Parallel Attention (EPA) module to emphasize informative encoder features before fusion. In the decoder, the architecture relies solely on the CSPF module, without combining it with CFGC, thereby reducing architectural complexity and focusing the decoder on progressive channel-wise feature fusion. The ablation results confirm that these modifications collectively improve feature representation, maintain segmentation accuracy, and significantly reduce parameter count and computational cost compared to the original X-UNet design.

The proposed CFCSE-Net achieves competitive performance with significantly fewer parameters and lower GFLOPs, as shown in Table 8. On the Kvasir-SEG dataset with an input resolution of 256 × 256, the model attains an IoU of 79.78% ± 1.99, a DSC of 87.21% ± 1.72, and an accuracy of 96.70% ± 0.59, indicating stable performance across different data splits. In addition, CFCSE-Net delivers the highest IoU and accuracy among the compared methods while maintaining a lightweight design with only 0.90M parameters and 3.24 GFLOPs. Although its DSC is slightly lower than that of X-UNet [14] (87.41 ± 1.62), the proposed model remains substantially more efficient than X-UNet (5.94M parameters, 10.72 GFLOPs). Despite the variability in polyp appearance, the model maintains consistent segmentation performance, offering an efficient alternative to more computationally intensive methods for real-time, resource-constrained applications.

For the CVC-ClinicDB dataset, Table 9 reports the performance of the model. The model achieves an IoU of 88.11% ± 0.86, a DSC of 93.42% ± 0.55, and an accuracy of 99.04% ± 0.09, demonstrating both high segmentation accuracy and stable performance across different data splits. The clearer texture and higher contrast in this dataset allow the network to capture object boundaries with better precision. Although its DSC remains slightly below the score of X-UNet [14],

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i1.1371

Table 7. Predicted outputs across all four datasets.

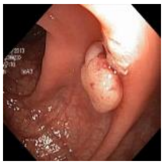
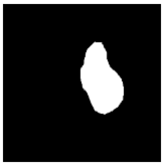
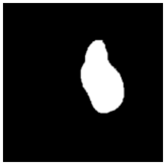
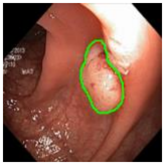
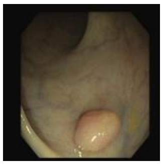
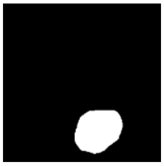
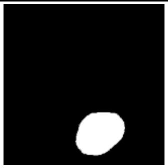
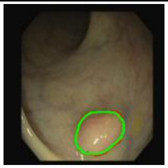
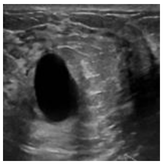
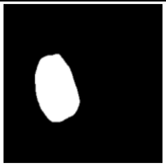
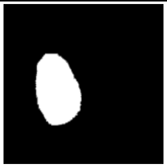
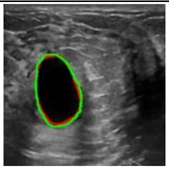
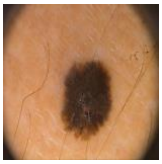
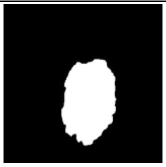
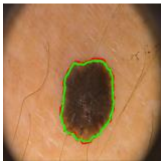
Dataset	Input Image	Ground Truth	Prediction	Overlay	IoU/DSC
Kvasir-SEG					96.20%/98.06%
CVC-ClinicDB					96.12%/98.02%
BUSI					90.71%/95.13%
PH^2					94.97%/97.42%

Table 8. Quantitative comparisons of various methods on the Kvasir-SEG dataset.

Method	Kvasir-SEG			Params	GFLOPs
	IoU/%	DSC/%	Acc/%		
U-Net [8]	75.55 ± 1.98	85.27 ± 1.59	95.63 ± 0.41	7.77M	13.75G
U-Net++ [9]	76.71 ± 2.22	86.07 ± 1.71	95.91 ± 0.37	9.16M	34.65G
AttUNet [27]	77.06 ± 1.71	86.27 ± 1.25	95.92 ± 0.31	8.73M	16.74G
TransUNet [28]	77.76 ± 1.43	86.84 ± 1.05	96.08 ± 0.21	105.32M	32.14G
UCTransUNet [29]	77.50 ± 2.51	86.59 ± 1.95	95.99 ± 0.55	66.49M	43.01G
CANet [15]	77.30 ± 0.52	86.31 ± 0.54	95.96 ± 0.10	24.12M	25.32G
TransAttUNet [30]	78.45 ± 2.23	87.26 ± 1.64	96.21 ± 0.45	25.97M	88.57G
BRAUNet++ [12]	75.79 ± 2.02	85.50 ± 1.61	95.66 ± 0.32	62.63M	22.33G
FDFUNet [31]	78.37 ± 2.38	87.17 ± 1.78	96.19 ± 0.53	20.95M	10.48G
X-UNet [14]	78.65 ± 2.19	**87.41 ± 1.62**	96.24 ± 0.34	5.94M	10.72G
Proposed Method	**79.78 ± 1.99**	87.21 ± 1.72	**96.70 ± 0.59**	**0.90M**	**3.24G**

the approach maintains high IoU and accuracy and uses a smaller number of parameters with lower computational cost. These results show that the method keeps high efficiency and adapts well to different polyp segmentation conditions. The results on the BUSI dataset are summarized in Table 10. Compared to other methods, BUSI presents a more challenging segmentation task due to the inherent characteristics of breast ultrasound images, including strong speckle noise, low contrast, heterogeneous tissue structures, and irregular lesion boundaries. These factors complicate accurate boundary localization and contribute to the lower IoU and DSC values observed on this dataset. Under these challenging conditions, the proposed CFCSE-Net achieves an IoU of 69.33% ± 2.66, a DSC of 78.80% ± 2.65, and an accuracy of 96.30% ± 0.51, while maintaining a lightweight design with only 0.90M parameters and 3.24 GFLOPs. Although these results are slightly lower than X-UNet [14] (69.57% ± 1.30 IoU, 80.34% ± 1.53 DSC, and 96.64% ± 0.21 accuracy), CFCSE-Net remains competitive while being substantially more efficient than X-UNet (5.94M parameters, 10.72 GFLOPs). This performance gap indicates that breast ultrasound segmentation may benefit from additional domain-specific enhancements, which remain a promising direction for future work.

The results on the PH^2 dataset are summarized in Table 11. Our model obtains the best scores for all evaluation metrics, with an IoU of 92.27% ± 0.52, a

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

Table 9. Quantitative comparisons of various methods on the CVC-ClinicDB dataset.

Method	CVC-ClinicDB			Params	GFLOPs
	IoU/%	DSC/%	Acc/%		
U-Net [8]	84.88 ± 0.75	91.56 ± 0.50	98.49 ± 0.12	7.77M	13.75G
U-Net++ [9]	86.48 ± 0.13	92.60 ± 0.09	98.67 ± 0.04	9.16M	34.65G
AttUNet [27]	86.22 ± 0.46	92.45 ± 0.29	98.65 ± 0.02	8.73M	16.74G
TransUNet [28]	86.95 ± 0.65	92.91 ± 0.40	98.73 ± 0.05	105.32M	32.14G
UCTransUNet [29]	87.47 ± 0.50	93.22 ± 0.31	98.77 ± 0.12	66.49M	43.01G
CANet [15]	86.88 ± 0.75	92.86 ± 0.48	98.72 ± 0.13	24.12M	25.32G
TransAttUNet [30]	87.66 ± 0.36	93.33 ± 0.25	98.79 ± 0.10	25.97M	88.57G
BRAUNet++ [12]	86.12 ± 0.37	92.40 ± 0.26	98.62 ± 0.08	62.63M	22.33G
FDFUNet [31]	87.81 ± 0.14	93.42 ± 0.09	98.84 ± 0.07	20.95M	10.48G
X-UNet [14]	87.90 ± 0.14	**93.49 ± 0.07**	98.84 ± 0.06	5.94M	10.72G
Proposed Method	**88.11 ± 0.86**	93.42 ± 0.55	**99.04 ± 0.09**	**0.90M**	**3.24G**

Table 10. Quantitative comparisons of various methods on the BUSI dataset.

Method	BUSI			Params	GFLOPs
	IoU/%	DSC/%	Acc/%		
U-Net [8]	63.25 ± 1.07	75.42 ± 1.04	96.19 ± 0.30	7.77M	13.75G
U-Net++ [9]	64.40 ± 0.92	76.57 ± 0.88	96.27 ± 0.28	9.16M	34.65G
AttUNet [27]	65.27 ± 1.46	77.05 ± 1.53	96.43 ± 0.26	8.73M	16.74G
TransUNet [28]	66.05 ± 2.23	77.61 ± 2.34	96.02 ± 0.06	105.32M	32.14G
UCTransUNet [29]	67.17 ± 1.20	78.47 ± 1.64	96.26 ± 0.26	66.49M	43.01G
CANet [15]	68.52 ± 1.20	79.49 ± 1.60	96.61 ± 0.22	24.12M	25.32G
TransAttUNet [30]	67.56 ± 1.41	78.99 ± 1.43	96.42 ± 0.25	25.97M	88.57G
BRAUNet++ [12]	66.67 ± 1.37	78.21 ± 1.54	96.20 ± 0.34	62.63M	22.33G
FDFUNet [31]	68.12 ± 0.36	79.28 ± 0.65	96.48 ± 0.28	20.95M	10.48G
X-UNet [14]	**69.57 ± 1.30**	**80.34 ± 1.53**	**96.64 ± 0.21**	5.94M	10.72G
Proposed Method	69.33 ± 2.66	78.80 ± 2.65	96.30 ± 0.51	**0.90M**	**3.24G**

DSC of 95.92% ± 0.30, and an accuracy of 98.06% ± 0.16. The low standard deviation values indicate highly consistent predictions across different data splits, demonstrating the robustness of the proposed model. Even with an input resolution of 320 × 320 pixels, the network stays efficient and uses only 0.90M parameters and 5.07 GFLOPs. Several transformer-based models perform poorly on this dataset because the training set contains only a small number of images. In contrast, the proposed approach maintains strong and stable results, indicating robust performance under data-constrained conditions. Across all four datasets, the proposed method uses the fewest parameters and GFLOPs among the compared models. It achieves the highest IoU and accuracy on Kvasir-SEG and CVC-ClinicDB, and it achieves the best overall performance on PH^2. However, it performs slightly below X-UNet on BUSI for IoU, DSC, and accuracy. BUSI images contain strong speckle noise and low contrast, which can blur lesion boundaries and reduce overlap metrics. This gap suggests that BUSI segmentation needs additional ultrasound-specific enhancements to better handle noise and irregular boundaries.

Although CFCSE-Net shows competitive accuracy with substantially fewer parameters and GFLOPs than many existing methods, several limitations should be noted. The model uses only four benchmark datasets, so it does not demonstrate generalization to other imaging modalities or to real clinical data. Some datasets, especially PH^2, contain limited samples, so the model may show different robustness on larger cohorts even with augmentation. The model also trains and tests only on individual 2D images after acquisition, so it does not evaluate real-time video conditions such as motion blur, frame-to-frame variation, and latency constraints. These limitations affect the interpretation of the results, because the reported metrics may change in broader clinical settings and continuous workflows. Therefore, the current findings mainly support performance on the selected benchmarks and offline 2D segmentation.

Despite these limitations, the balance between accuracy and efficiency shows the practical relevance of the proposed architecture for deployment in resource-constrained environments. With approximately 0.90 million parameters and low computational cost (3.24 GFLOPs for 256 × 256 inputs and 5.07 GFLOPs for 320 × 320 inputs), CFCSE-Net supports computer-aided diagnosis (CAD) systems on edge devices and low-power clinical workstations,

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i1.1371

Table 11. Quantitative comparisons of various methods on the PH^2 dataset.

Method	PH^2			Params	GFLOPs
	IoU/%	DSC/%	Acc/%		
U-Net [8]	90.15 ± 0.46	94.78 ± 0.26	96.47 ± 0.16	7.77M	21.49G
U-Net++ [9]	90.46 ± 0.68	94.96 ± 0.39	96.60 ± 0.08	9.16M	54.14G
AttUNet [27]	90.09 ± 0.56	94.74 ± 0.31	96.40 ± 0.16	8.73M	26.16G
TransUNet [28]	91.21 ± 0.05	95.38 ± 0.02	96.82 ± 0.02	105.32M	50.21G
UCTransUNet [29]	90.44 ± 0.85	94.94 ± 0.49	96.48 ± 0.66	66.49M	67.19G
CANet [15]	91.03 ± 0.54	95.28 ± 0.30	96.87 ± 0.24	24.12M	39.56G
TransAttUNet [30]	90.27 ± 0.22	94.86 ± 0.12	96.57 ± 0.18	25.97M	138.38G
BRAUNet++ [12]	90.69 ± 0.26	95.09 ± 0.14	96.67 ± 0.07	62.63M	34.89G
FDFUNet [31]	91.38 ± 0.33	95.48 ± 0.17	96.98 ± 0.26	23.24M	16.38G
X-UNet [14]	91.49 ± 0.09	95.55 ± 0.05	97.09 ± 0.13	5.94M	16.74G
Proposed Method	**92.27 ± 0.52**	**95.92 ± 0.30**	**98.06 ± 0.16**	**0.90M**	**5.07G**

where memory usage, inference latency, and energy efficiency are critical. This design enables faster inference and supports near-real-time feedback in clinical workflows such as endoscopy and ultrasound examination. Future work should further evaluate the approach on additional imaging modalities and real-time clinical scenarios, including extensions to volumetric 3D data through separable or pseudo-3D convolutional operations and selective attention along spatial or depth dimensions, as well as video-based segmentation using lightweight temporal feature reuse or frame-to-frame refinement strategies, while preserving the core lightweight design of CFCSE-Net.

V. Conclusion

This study proposes CFCSE-Net, a U-Net-based architecture with a modified Collaborative Fusion with Global Context-Aware (CFGC) module with added Ghost Modules in the encoder, Cross Split-Channel Progressive Fusion (CSPF) modules in the decoder, and Enhanced Parallel Attention (EPA) along the skip connections. The goal is to improve segmentation accuracy while keeping computational cost low for use in resource-constrained CAD systems. Extensive experiments conducted on four benchmark datasets using multiple random data splits demonstrate both strong performance and high stability. The model achieves mean IoU, DSC, and accuracy values of 79.78% ± 1.99, 87.21% ± 1.72, and 96.70% ± 0.59 on Kvasir-SEG; 88.11% ± 0.86, 93.42% ± 0.55, and 99.04% ± 0.09 on CVC-ClinicDB; 69.33% ± 2.66, 78.80% ± 2.65, and 96.30% ± 0.51 on BUSI; and 92.27% ± 0.52, 95.92% ± 0.30, and 98.06% ± 0.16 on PH2. The relatively low standard deviations across all datasets indicate consistent performance across different data splits and confirm the robustness of the proposed approach. The model maintains a lightweight structure with about 0.90 million parameters and a computational cost of 3.24 GFLOPs for 256 × 256 inputs and 5.07 GFLOPs for 320 × 320 inputs. These results show that the combination of multi-scale feature extraction and attention mechanism improves segmentation performance while preserving efficiency. In the future, CFCSE-Net may be extended to other imaging modalities such as CT, MRI, and 3D volumetric data, adapt the model for real-time video segmentation in endoscopic procedures, explore semi-supervised or self-supervised learning to use unlabeled data, and apply model compression and quantization so that deployment on edge devices in clinical settings becomes more practical.

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 2023 under the research grant Hibah Penelitian Fundamental (PF-UNS) with the contract 228/UN27.22/PT.01.03/2023.

Data Availability

The datasets used in this study are publicly accessible. We obtained the data from Kaggle repositories associated with the works of Jha et al. [16], Bernal et al. [17], Al-Dhabyani et al. [18], and Mendonca et al. [19]. The Methods section describes the full preprocessing

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

pipeline and evaluation protocol to ensure reproducibility.

Author Contribution

Alfath Roziq Widhayaka 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 uses four open-source medical image datasets (Kvasir-SEG [16], CVC-ClinicDB [17], BUSI [18], and PH2 [19]) for segmentation. These datasets are publicly available and provide anonymized medical images. Therefore, this study did not require additional ethical approval.

Consent for Publication

All participants gave consent for publication.

Competing Interest

The authors declare no competing interests.

References

[1] J. Zhang *et al.*, "Advances in attention mechanisms for medical image segmentation," *Computer Science Review*, vol. 56, p. 100721, May 2025, doi: 10.1016/j.cosrev.2024.100721.

[2] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, "Medical Image Segmentation based on U-Net: A Review," *jist*, vol. 64, no. 2, pp. 020508-1-020508–12, Mar. 2020, doi: 10.2352/J.ImagingSci.Technol.2020.64.2.020508.

[3] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, and D. Visvikis, "Current and Emerging Trends in Medical Image Segmentation With Deep Learning," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 6, pp. 545–569, July 2023, doi: 10.1109/TRPMS.2023.3265863.

[4] X. Shu, J. Wang, A. Zhang, J. Shi, and X.-J. Wu, "CSCA U-Net: A channel and space compound attention CNN for medical image segmentation," *Artificial Intelligence in Medicine*, vol. 150, p. 102800, Apr. 2024, doi: 10.1016/j.artmed.2024.102800.

[5] Y. Zhang, Q. Liao, L. Ding, and J. Zhang, "Bridging 2D and 3D segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5D solutions," *Computerized Medical Imaging and Graphics*, vol. 99, p. 102088, July 2022, doi: 10.1016/j.compmedimag.2022.102088.

[6] N. M. Ali, S. S. Oyelere, N. Jitani, R. Sarmah, and S. Andrew, "Hybrid intelligence in medical image segmentation," *Sci Rep*, vol. 15, no. 1, p. 41200, Nov. 2025, doi: 10.1038/s41598-025-24990-w.

[7] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, *arXiv*. doi: 10.48550/ARXIV.1505.04597.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," 2018, *arXiv*. doi: 10.48550/ARXIV.1807.10165.

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

[11] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More Features From Cheap Operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, June 2020, pp. 1577–1586. doi: 10.1109/CVPR42600.2020.00165.

[12] L. Lan, P. Cai, L. Jiang, X. Liu, Y. Li, and Y. Zhang, "BRAU-Net++: U-Shaped Hybrid CNN-Transformer Network for Medical Image Segmentation," 2024, *arXiv*. doi: 10.48550/ARXIV.2401.00722.

[13] M. A. Fathur Rohman, H. Prasetyo, E. P. Yudha, and C.-H. Hsia, "Improving Accuracy and Efficiency of Medical Image Segmentation Using One-Point-Five U-Net Architecture with Integrated Attention and Multi-Scale Mechanisms," *j.electron.electromedical.eng.med.inform*, vol. 7, no. 3, pp. 869–880, July 2025, doi: 10.35882/jeeemi.v7i3.949.

[14] S. Xu *et al.*, "X-UNet:A novel global context-aware collaborative fusion U-shaped network with progressive feature fusion of codec for medical image segmentation," *Neural Networks*, vol. 192, p. 107943, Dec. 2025, doi: 10.1016/j.neunet.2025.107943.

[15] X. Xie *et al.*, "CANet: Context aware network with dual-stream pyramid for medical image segmentation," *Biomedical Signal Processing*

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

and Control, vol. 81, p. 104437, Mar. 2023, doi: 10.1016/j.bspc.2022.104437.

[16] D. Jha *et al.*, "Kvasir-SEG: A Segmented Polyp Dataset," 2019, *arXiv*. doi: 10.48550/ARXIV.1911.07069.

[17] 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," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, July 2015, doi: 10.1016/j.compmedimag.2015.02.007.

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

[19] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2 - A dermoscopic image database for research and benchmarking," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka: IEEE, July 2013, pp. 5437–5440. doi: 10.1109/EMBC.2013.6610779.

[20] M. A. Fathur Rohman, H. Prasetyo, H. M. Akbar, and A. D. Afan Firdaus, "ACMU-Net: An Efficient Architecture Based on ConvMixer and Attention Mechanism for Colorectal Polyp Segmentation," in *2024 IEEE International Conference on Smart Mechatronics (ICSMech)*, Yogyakarta, Indonesia: IEEE, Nov. 2024, pp. 279–284. doi: 10.1109/ICSMech62936.2024.10812309.

[21] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artif Intell Rev*, vol. 56, no. 11, pp. 12561–12605, Nov. 2023, doi: 10.1007/s10462-023-10453-z.

[22] L. Lu, Q. Xiong, D. Chu, and B. Xu, "MixDehazeNet : Mix Structure Block For Image Dehazing Network," 2023, *arXiv*. doi: 10.48550/ARXIV.2305.17654.

[23] R. Andonie, "Hyperparameter optimization in learning systems," *J Membr Comput*, vol. 1, no. 4, pp. 279–291, Dec. 2019, doi: 10.1007/s41965-019-00023-0.

[24] Y. Yuan and Y. Cheng, "Medical image segmentation with UNet-based multi-scale context fusion," *Sci Rep*, vol. 14, no. 1, p. 15687, Oct. 2024, doi: 10.1038/s41598-024-66585-x.

[25] H. Al Jowair, M. Alsulaiman, and G. Muhammad, "Multi parallel U-net encoder network for effective polyp image segmentation," *Image and Vision Computing*, vol. 137, p. 104767, Sept. 2023, doi: 10.1016/j.imavis.2023.104767.

[26] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019, doi: 10.48550/ARXIV.1905.11946.

[27] S. Wang, L. Li, and X. Zhuang, "AttU-NET: Attention U-Net for Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, vol. 12963, A. Crimi and S. Bakas, Eds., in Lecture Notes in Computer Science, vol. 12963. , Cham: Springer International Publishing, 2022, pp. 302–311. doi: 10.1007/978-3-031-09002-8_27.

[28] J. Chen *et al.*, "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, Oct. 2024, doi: 10.1016/j.media.2024.103280.

[29] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer," 2021, *arXiv*. doi: 10.48550/ARXIV.2109.04335.

[30] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation," 2021, *arXiv*. doi: 10.48550/ARXIV.2107.05274.

[31] Y. Chen, X. Zhang, L. Peng, Y. He, F. Sun, and H. Sun, "Medical image segmentation network based on multi-scale frequency domain filter," *Neural Networks*, vol. 175, p. 106280, July 2024, doi: 10.1016/j.neunet.2024.106280.

Author Biography

Alfath Roziq Widhayaka is a Computer Science student at Universitas Sebelas Maret (UNS) with a strong interest in deep learning and image processing. He actively explores advanced topics in neural networks and medical image analysis, driven by a passion for understanding and developing innovative methods in his field. He achieved third place at the national-level GEMASTIK competition in Division VII Scientific Writing with the proposed work titled DD-GMANet. Throughout his academic journey, he continuously seeks new knowledge and challenges to strengthen his expertise in artificial intelligence and its real-world applications. He can be contacted at email: alfathroziq94@student.uns.ac.id.

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371

Heri Prasetyo received the doctoral degree from the Department of Electrical Engineering, National Taiwan University of Science and Technology (NTUST), Taiwan, in 2015. He was awarded the Best Dissertation Award from the Taiwan Association for Consumer Electronics (TACE) in 2015, and has received multiple Best Paper Awards including 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), the International Conference on Artificial Intelligence's 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 email: heri.prasetyo@staff.uns.ac.id.

Manuscript received 8 September 2025; Revised 10 December 2025; Accepted 20 December 2025; Available online 9 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1371