

Reduction of decoder convolutional blocks in Attention Gate U-Net++ to enhance computational efficiency in automatic HR-CTV delineation on CT-based cervical cancer brachytherapy images

Thinnagit Srihot¹, Chatrawut Pattaweerakul², Kamonchanok Nobphuek², Titipong Kaewlek^{1,3*}

¹Medical Physics Program, Department of Radiological Technology, Faculty of Allied Health Sciences, Naresuan University, Mueang District, Phitsanulok Province, Thailand.

²Department of Radiology, Faculty of Medicine, Naresuan University, Mueang District, Phitsanulok Province, Thailand.

³Interdisciplinary Health and Data Sciences Research Unit, Faculty of Allied Health Sciences, Naresuan University, Phitsanulok Province, Thailand.

ARTICLE INFO

Article history:

Received 15 April 2026

Accepted as revised 21 July 2026

Available online 27 July 2026

Keywords:

Cervical cancer, U-Net, U-Net++,
attention U-Net++, brachytherapy

ABSTRACT

Background: The contour of the high-risk clinical target volume (HR-CTV) is important in computed tomography (CT)-based cervical cancer brachytherapy to ensure adequate target tumor coverage while sparing radiation exposure to organs at risk (OARs). Although Attention U-Net++ has demonstrated improved segmentation performance, its complex decoder structure substantially increases computational cost and training time.

Objectives: This study aimed to develop a computationally efficient modified Attention U-Net++ architecture by reducing decoder depth while preserving segmentation performance for HR-CTV delineation.

Materials and methods: A retrospective dataset comprising 3,323 CT images from 56 patients with cervical cancer was analyzed. The proposed model eliminates the deepest decoder level to reduce redundant feature propagation and overall model complexity. Segmentation performance was evaluated using the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), 95th percentile Hausdorff Distance (HD95), Accuracy, Precision, and Recall. Computational efficiency was assessed by the number of trainable parameters, training time, and inference time, and was compared with the baseline Attention U-Net++ architecture and related models.

Results: The proposed Attention U-Net++ DL3 reduced the number of trainable parameters from 37.30 million to 8.86 million and shortened training time from 52.27 minutes to 33.11 minutes, while maintaining a comparable inference time (36.90 seconds). Geometric evaluation demonstrated a DSC of 0.9076 and IoU of 0.8376, marginally outperforming the baseline Attention U-Net++ (DSC 0.9057; IoU 0.8338). The HD95 value was 1.3959, which is comparable to existing architectures. Pixel-level classification performance remained robust, with an accuracy of 0.9996, precision of 0.9046, and recall of 0.9298. Overall, the proposed model achieved competitive segmentation performance with substantially reduced model complexity.

Conclusion: Reducing decoder convolutional blocks in Attention U-Net++ significantly improves computational efficiency while preserving segmentation accuracy. The proposed lightweight architecture provides an optimal balance between performance and resource utilization, supporting its practical implementation in CT-based cervical cancer brachytherapy, particularly in resource-limited clinical settings.

* Corresponding contributor.

Author's Address: Medical Physics Program,
Department of Radiological Technology,
Faculty of Allied Health Sciences, Naresuan
University, Mueang District, Phitsanulok
Province, Thailand.

E-mail address: titipongk@nu.ac.th

doi: 10.66285/JAMS.2026.104

E-ISSN: 2539-6056

Introduction

Gynecological cancers arise within the female reproductive organs and primarily include cervical, ovarian, and endometrial cancers. Among these, cervical cancer remains one of the leading causes of cancer-related mortality in women worldwide. Data from the Surveillance, Epidemiology, and End Results (SEER) program indicate that cervical cancer accounted for approximately 342,000 deaths between 1975 and 2020, representing the highest mortality rate among gynecological malignancies during this period.¹ Cervical cancer predominantly affects women aged 30-70 years and is strongly associated with persistent infection with high-risk human papillomavirus (HPV), which can induce precancerous cellular abnormalities and ultimately lead to malignant transformation of cervical epithelial cells.²

Brachytherapy is an essential treatment modality for cervical cancer, particularly in locally advanced disease. This technique delivers high-dose radiation directly to the tumor using intracavitary applicators inserted into the uterus, enabling precise target coverage while minimizing radiation exposure to surrounding healthy tissues.³ Current clinical practice follows the recommendations of the GEC-ESTRO Working Group IV, which advocates the use of magnetic resonance imaging (MRI) as the primary imaging modality for brachytherapy planning due to its superior soft-tissue contrast and improved accuracy in delineating target volumes and organs at risk (OARs).⁴ However, MRI is not universally available in all clinical settings. In such cases, the IBS-GEC ESTRO-ABS guidelines recommend computed tomography (CT) as an alternative, provided that a pre-treatment MRI is performed to guide CT-based treatment planning throughout the brachytherapy course.⁵

Despite advances in imaging and treatment techniques, CT-based cervical cancer brachytherapy remains challenging. Accurate delineation of the high-risk clinical target volume (HR-CTV) is critical because high radiation doses must be confined to the tumor while sparing adjacent organs, such as the bladder and rectum. Manual contouring is time-consuming and operator-dependent; furthermore, patients are often required to remain immobilized with an intracavitary applicator in place for prolonged periods, which may cause discomfort and reduce treatment tolerance.⁶

To address these challenges, deep learning methods have increasingly been applied to medical image segmentation. Ronneberger *et al.* introduced the U-Net architecture, which demonstrated strong performance in biomedical image segmentation tasks.⁷ However, subsequent studies reported that the performance of the conventional U-Net may degrade when segmenting small or low-contrast anatomical structures due to the inclusion of irrelevant background features.⁸ To mitigate this limitation, attention mechanisms were introduced to enable selective focus on salient regions.⁹ In parallel,

Zhou *et al.* proposed U-Net++, which enhances feature representation through densely connected convolutional blocks and deep supervision, thereby improving training stability and segmentation performance.¹⁰ More recently, attention mechanisms have been integrated into U-Net++ skip connections to further refine feature selection.¹¹ However, these modifications substantially increase model complexity, resulting in higher computational costs and longer training times.

Although Attention U-Net++ improves segmentation accuracy, its high number of trainable parameters may limit practical implementation in clinical settings with constrained computational resources.¹² To date, limited studies have systematically investigated architectural simplification strategies for Attention U-Net++ in CT-based cervical cancer brachytherapy segmentation. Therefore, this study aimed to develop and evaluate a lightweight modified Attention U-Net++ architecture by reducing decoder depth while maintaining segmentation performance.¹³

Materials and methods

This study was conducted to evaluate the proposed segmentation model for CT-based cervical cancer brachytherapy. Ethical approval was obtained from the Institutional Review Board prior to data collection. The study utilized medical imaging data from patients treated between 2022 and 2026. All patient data were anonymized prior to analysis to ensure confidentiality and compliance with ethical standards.

Patient selection

A total of 56 patients diagnosed with cervical cancer, aged between 30-70 years, were included in this study. All patients were clinically staged according to the International Federation of Gynecology and Obstetrics (FIGO) staging system (IA-IIIIC). Each patient underwent high-dose-rate (HDR) intracavitary brachytherapy using the Fletcher-Suit applicator system, consisting of tandem and ovoid components (3-4 channels). The radioactive source used for treatment was iridium-192 (¹⁹²Ir). A prescribed dose of 7 Gy per fraction was delivered for 1-4 fractions depending on clinical indications. Prior to each brachytherapy session, CT simulation was performed with a standardized bladder filling protocol of 50 mL sterile water to maintain consistent organ positioning. CT images were acquired using the Philips Brilliance Big Bore system with a slice thickness of 2 mm. After-loader brachytherapy treatment was delivered using the GammaMed Plus iX system. All imaging data were stored in Digital Imaging and Communications in Medicine (DICOM) format (Table 1). Radiation therapy structures (RT-Structure), including target volumes and organs at risk, were manually delineated by a board-certified radiation oncologist with more than 10 years of clinical experience.

Table 1. Patient characteristics.

Characteristics	
Age (years)	30-70
Retrospective study selection	2022-2026
Radiation source	Iridium-192
Prescription dose per fraction (Gy)	7 Gy x 1-4 fractions
Applicator	Fletcher suit
FIGO clinical stage	IA-IIIC
No. of patients	56
CT slice thickness (mm)	2

Note: CT: computed tomography, FIGO: International Federation of Gynecology and Obstetrics.

Patient dataset preparation

A total of 3,323 CT images were collected from 56 patients. On average, each patient underwent approximately four CT simulation sessions (fractions) during brachytherapy treatment, resulting in multiple CT image sets per patient. To prevent data leakage and ensure an unbiased evaluation of model performance, the dataset was partitioned on a patient-wise basis, such that all CT images from the same patient, including all treatment fractions, were assigned exclusively to a single subset.

The dataset was divided into three subsets for deep learning model development: the training set consisted of 2,684 CT images from 37 patients, the validation set comprised 299 CT images from

8 patients, and the testing set included 340 CT images from 11 patients. This patient-wise splitting strategy ensured that no CT images or treatment fractions from the same patient were shared across the training, validation, and testing subsets, thereby minimizing the risk of information leakage and providing a more reliable assessment of the model's generalization performance.

Ground-truth segmentation masks were generated from Digital Imaging and Communications in Medicine (DICOM) datasets comprising CT image files and their corresponding Radiation Therapy Structure Set (RT-STRUCT) files. All datasets were imported into 3D Slicer (version 5.8.1) for visualization and contour processing (Figure 1).

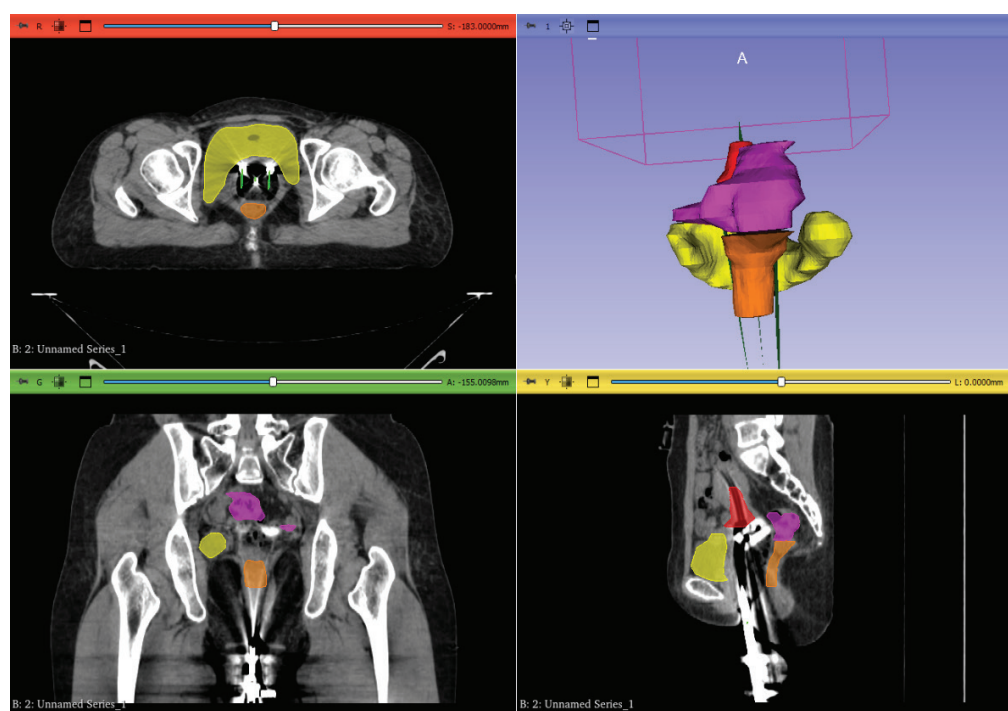


Figure 1. Visualization of pelvic anatomical structures, highlighting the delineated HR-CTV and organs at risk (OARs) using 3D Slicer software (version 5.8.1) derived from DICOM CT datasets.

The high-risk clinical target volume (HR-CTV) was derived from RTSTRUCT delineations and processed using 3D Slicer (version 5.8.1) to generate binary segmentation masks. The segmented volumes were exported in Neuroimaging Informatics Technology Initiative (NIfTI; .nii) format and subsequently imported into Google Colab for further processing. Both CT images and corresponding segmentation masks were resized from 512×512 pixels to 256×256 pixels to enhance computational efficiency.¹⁴ The processed images were saved in Portable Network Graphics (PNG) format using automated file naming and indexing procedures to maintain accurate CT-mask pairing. Finally, all images were converted into NumPy array (.npy) format for model training. Data augmentation techniques, including rotation, shearing, horizontal flipping, and vertical flipping, were applied to the training and validation sets to increase data variability and mitigate overfitting.¹⁵ The test set remained unaltered to ensure an unbiased performance evaluation.

The Attention Gate U-Net++ architecture

The integration of attention mechanisms into U-Net++ enhances the network's ability to focus on clinically relevant regions while suppressing irrelevant background information, particularly in low-contrast images and small target structures.¹¹ In the Attention

U-Net++, Attention Gates (AGs) are embedded within skip connections, where encoder feature maps are modulated by gating signals from deeper decoder layers to generate spatial attention coefficients prior to feature fusion. In this study, a modified Attention U-Net++ (AU-Net++) architecture was implemented using TensorFlow. The network accepts two-dimensional CT slices with an input size of 256×256×1 and follows an encoder-decoder structure with nested dense skip pathways. The encoder consists of four downsampling stages and a bottleneck layer, with the number of convolutional filters progressively increasing through 64, 128, 256, 512, and 1,024. Each stage contains two 3×3 convolutional layers with rectified linear unit (ReLU) activation, followed by 2×2 max pooling for spatial downsampling.

The decoder incorporates multi-level nested skip connections, with attention gates applied across multiple levels to refine encoder feature maps before concatenation. Each attention gate employs 1×1 convolutions with ReLU activation, followed by sigmoid activation to generate attention maps. These attention maps are multiplied elementwise with the corresponding encoder features to enhance relevant spatial regions. Finally, the output layer applies a 1×1 convolution with sigmoid activation to generate a binary segmentation probability map (Figure 2).

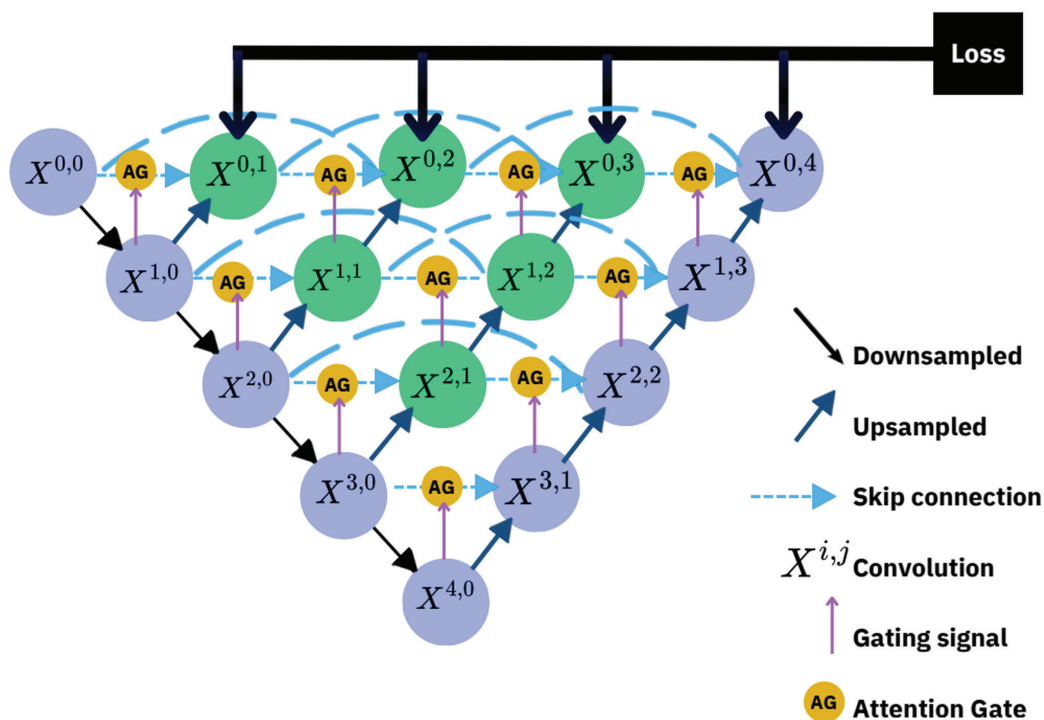


Figure 2. Architecture of the Attention Gate U-Net++.

The deep supervision of the Attention Gate U-Net++ architecture

Deep supervision facilitates stable gradient propagation and improves training convergence by applying auxiliary loss functions to intermediate decoder outputs.¹¹ In the Attention U-Net++ architecture, segmentation outputs are generated from multiple decoder levels (L^1 , L^2 , and L^3), as illustrated in Figure 3. Each intermediate output is passed through a 1×1 convolution layer followed by a sigmoid activation to

produce a probability map. During training, the loss function is computed at each supervised level, and the final loss is defined as the weighted sum of all intermediate losses. This strategy encourages feature refinement at multiple semantic scales and mitigates the vanishing gradient problem in deeper architectures. During inference, only the final decoder output is used to generate the segmentation result, while intermediate outputs are utilized solely to guide the training process.

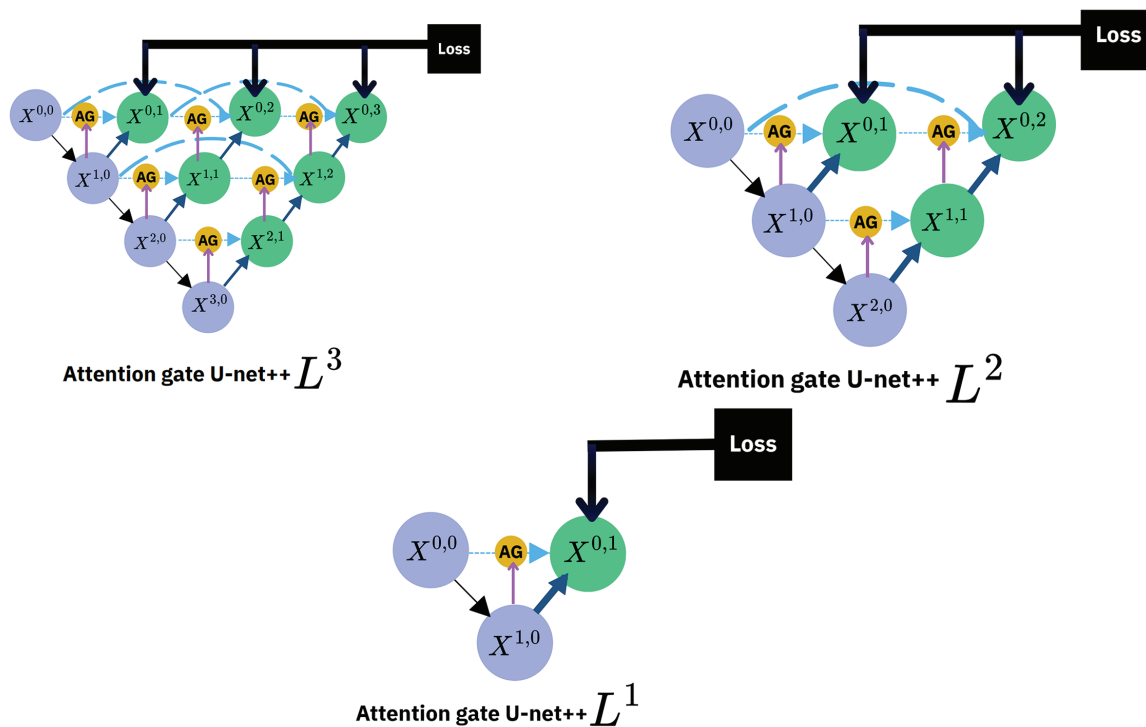


Figure 3. Deep supervision mechanism in the Attention Gate U-Net++.

The Attention Gate calculation process

The Attention Gate mechanism operates by jointly processing the upsampled decoder features and the corresponding encoder features from the skip connection.⁸⁻¹⁶ Both feature maps are first projected through separate 1×1 convolution layers to reduce and align their channel dimensions. The transformed features are then combined using element-wise summation and passed through a rectified linear unit (ReLU) activation function to introduce non-linearity.

Subsequently, a 1×1 convolution followed by a sigmoid activation function is applied to generate the attention coefficient (α), which ranges between 0 and 1. The resulting attention map is upsampled to match the spatial resolution of the encoder feature map and is then multiplied elementwise with the encoder features. This gating operation suppresses irrelevant background responses while enhancing salient features, thereby improving feature representation for subsequent decoding stages (Figure 4).

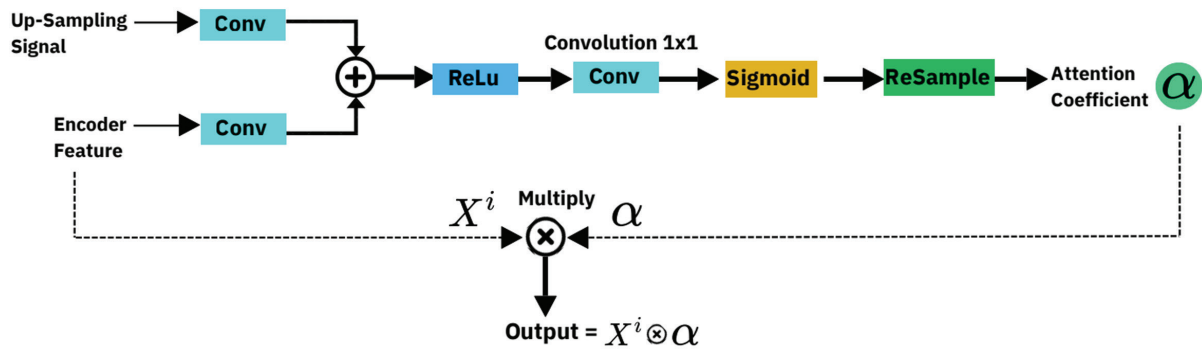


Figure 4. The Attention Gate calculation process.

The proposed network architecture

To reduce the computational burden while preserving segmentation performance, a modified Attention U-Net++ (AU-Net++) architecture was developed by progressively reducing decoder depth.¹⁷ As illustrated in Figure 5, three configurations were evaluated: AU-Net++ DL3 (removal of decoder level 4), AU-Net++ DL2 (removal of decoder level 3), and AU-Net++ DL1 (removal of decoder level 2). In deep encoder-decoder architectures, repeated downsampling during feature extraction progressively reduces spatial resolution, leading to the loss of fine-grained boundary information.¹⁸ Consequently, deeper decoder layers may contribute limited spatial detail due to the

accumulation of coarse semantic features and boundary blurring. This observation suggests that the deepest decoder levels may introduce redundant computations without substantially improving geometric precision. Based on this rationale, higher decoder levels were systematically removed to reduce model complexity and computational cost. By truncating the deepest decoding stages, the proposed architectures aim to retain sufficient semantic representation while minimizing unnecessary feature propagation.¹⁹ This structural simplification decreases the number of trainable parameters and training time, while potentially improving geometric accuracy by mitigating boundary over-smoothing in the final segmentation output.

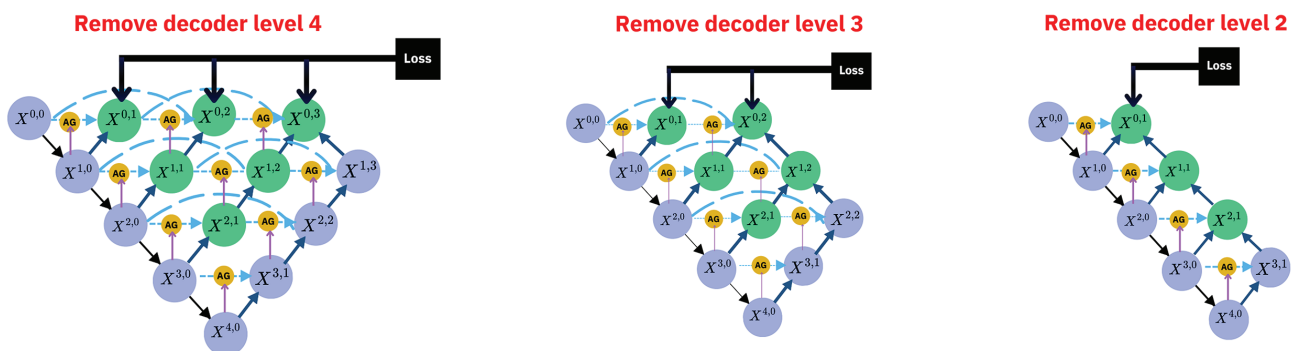


Figure 5. The AU-Net++ architectures with progressively reduced decoder depths: DL3 (removing Level 4), DL2 (removing Level 3), and DL1 (removing Level 2) to improve computational efficiency.

Model training

Model development and preliminary experiments were conducted on a laptop (MSI Katana; Intel® Core™ i7 processor; NVIDIA GeForce RTX 4060 GPU). For large-scale training and final model optimization, experiments were implemented in Python using the Keras API integrated within the TensorFlow framework and executed on the Google Colab platform utilizing an NVIDIA A100 GPU runtime. To determine the optimal training configuration, various hyperparameter settings were evaluated. The network was trained using three epoch settings (60, 80, and 100 epochs) and three

batch sizes (8, 16, and 32). Dice+Binary Cross-entropy loss is defined in Equation (1):

$$L_{BCE} = -[g \log(p) + (1 - g) \log(1 - p)] \quad (1)$$

where p denotes the predicted probability produced by the model after the sigmoid activation, and g represents the corresponding binary ground-truth label. The dice similarity coefficient (DSC) was employed to directly optimize segmentation overlap performance. The Adam optimizer was applied with an initial learning rate of 0.001. Training performance was monitored using

validation loss and DSC metrics to assess convergence behavior. Experimental results indicated that training with 60 epochs and a batch size of 32 achieved stable convergence without observable signs of overfitting, as reflected by consistent validation performance. Consequently, this configuration was selected as the optimal setting, providing stable training behavior, satisfactory segmentation performance, reduced training time, and efficient GPU resource utilization. Training time and inference time were quantitatively evaluated using Python's time library. Furthermore, the total number of trainable parameters was calculated to assess model complexity, particularly in relation to the reduced decoder depth.

Statistical analysis

To assess segmentation performance, both geometric accuracy metrics and pixel-level classification metrics were evaluated. Geometric accuracy was quantified using the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and the 95th percentile Hausdorff Distance (HD95). The DSC and IoU measure the spatial overlap between the predicted segmentation mask and the ground-truth annotation. The HD95 evaluates boundary agreement by computing the 95th percentile of the bidirectional surface distance between predicted and reference contours, thereby reducing sensitivity to extreme outlier points. These metrics are defined in Equations (2)-(4).

$$\text{DSC} = \frac{2|A \cap B|}{|A| + |B|} \quad (2)$$

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

$$\text{HD}_{95}(A, B) = \max \{P_{95}(\min_{b \in B}(a, b)), P_{95}(\min_{a \in A}(b, a))\} \quad (4)$$

where A and B denote the sets of boundary points for the ground-truth and predicted segmentation masks, respectively, d represents the Euclidean distance, and indicates the 95th percentile.

The model performance was further evaluated using accuracy, precision, and recall at the pixel level.

Accuracy represents the overall proportion of correctly classified pixels. Precision measures the proportion of correctly predicted positive pixels to all predicted positive pixels, while recall quantifies the proportion of true positive pixels correctly identified by the model. These metrics are formally defined in Equations (5)-(7).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

Specifically, true positive (TP) represents pixels correctly classified as the target region, true negative (TN) denotes pixels correctly classified as background, false positive (FP) signifies background pixels incorrectly classified as target, and false negative (FN) refers to target pixels incorrectly classified as background.

Results

Evaluation of model training

Comparison of the computational efficiency in model evaluation is demonstrated in Table 2. The baseline Attention U-Net++ (AU-Net++) model contains 37.30 million trainable parameters, with a training time of 52.27 minutes and an inference time of 36.97 seconds per test set. Following decoder depth reduction, the AU-Net++ DL3 reduced the number of parameters to 8.86 million and the training time to 33.11 minutes. The AU-Net++ DL2 further decreased the parameter count to 2.10 million, with a corresponding training time of 19.41 minutes. The most compact configuration, AU-Net++ DL1, achieved the smallest model size with only 0.41 million parameters and a training time of 9.26 minutes. A consistent reduction in inference time was also observed across models, with AU-Net++ DL1 demonstrating the fastest performance at 33.29 seconds per test set. These findings indicate that progressive decoder reduction substantially improves computational efficiency while maintaining practical inference speed.

Table 2. Comparison of computational efficiency in model evaluation.

Model	Parameters	Training time (minutes)	Inference time (seconds/test set)
Attention U-net++	37.30 M	52.27	36.9660
AU-Net++ DL3	8.86 M	33.11	36.8953
AU-Net++ DL2	2.10 M	19.41	35.2526
AU-Net++ DL1	0.41 M	09.26	33.2896

Note: M: millions.

Evaluation of model stability

Figure 6 illustrates the training and validation Dice loss curves for A: Attention Gate U-Net++, B: AU-Net++ DL3, C: AU-Net++ DL2, and D: AU-Net++ DL1. All models exhibit a rapid reduction in loss during the

initial epochs, followed by stable convergence. The training and validation curves remain closely aligned across all configurations, indicating stable learning behavior and the absence of significant overfitting.

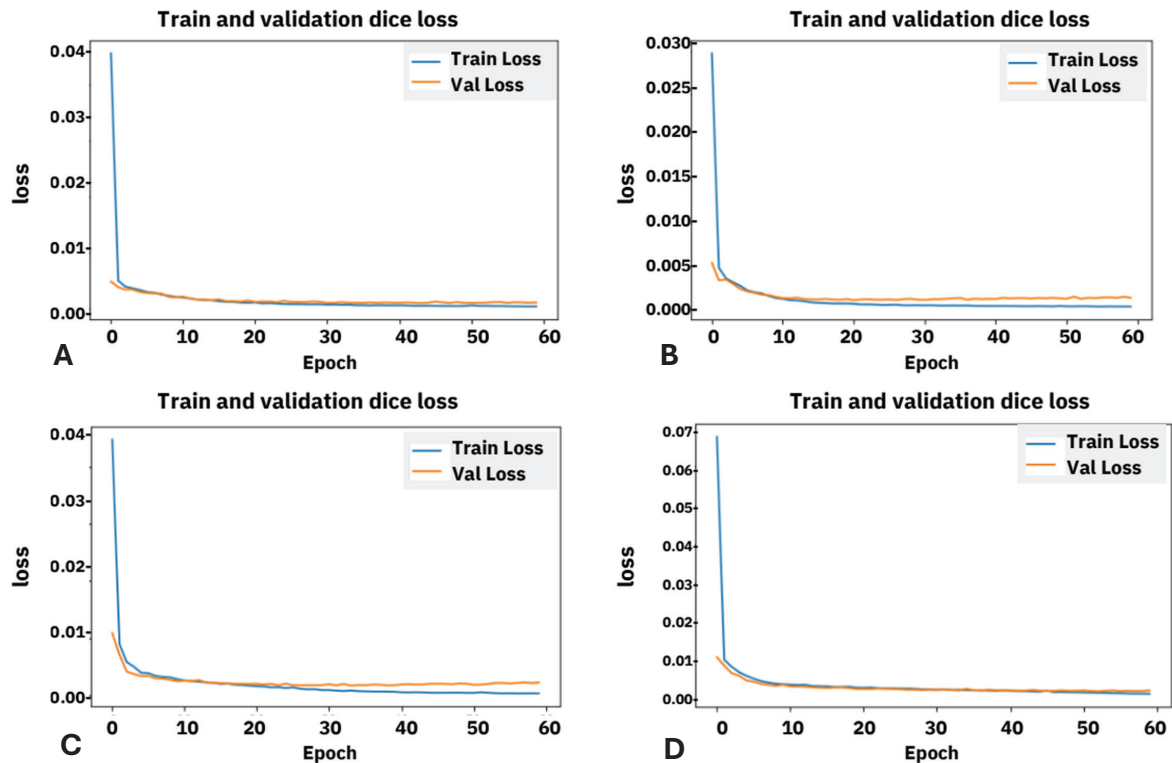


Figure 6. Training and validation Binary Cross-entropy Dice loss curves. A: Attention Gate U-Net++ model, B: AU-Net++ DL3, C: AU-Net++ DL2, D: AU-Net++ DL1.

Geometric and performance model evaluation

Table 3 demonstrates that the AU-Net++ DL3 model achieved the best overlap performance, with the highest DSC of 0.9076 and IoU of 0.8376, slightly outperforming Attention U-Net++ (DSC: 0.9057, IoU: 0.8338). Attention U-Net++ obtained the lowest HD95 (1.3279), while DL3 exhibited comparable boundary performance (1.3959). Both models achieved identical

accuracy (0.9996), with DL3 providing higher precision (0.9046) and Attention U-Net++ yielding higher recall (0.9382). Performance declined for DL2 and DL1, particularly regarding DSC, IoU, and HD95, although DL1 achieved the highest precision (0.9162). Overall, DL3 offers the optimal balance between accuracy and model simplification.

Table 3. Geometric and performance evaluation.

Model	DSC	IoU	HD95	Accuracy	Precision	Recall
Attention U-net++	0.9057	0.8338	1.3279	0.9996	0.8920	0.9382
AU-Net++ DL3	0.9076	0.8376	1.3959	0.9996	0.9046	0.9298
AU-Net++ DL2	0.8678	0.7783	1.6236	0.9995	0.8729	0.8928
AU-Net++ DL1	0.7514	0.6149	3.3098	0.9991	0.9162	0.6721

The results of boundary overlap analysis

Figure 7 presents the visual comparison between the ground-truth contour (solid red line) and the predicted contour (solid blue line) for Attention Gate U-Net++, AU-Net++ DL3, AU-Net++ DL2, and AU-Net++ DL1. The Attention Gate U-Net++ model demonstrates

overall strong boundary agreement, although slight deviations are observed in some peripheral regions. AU-Net++ DL3 exhibits highly precise alignment between the predicted and ground-truth contours, with enhanced consistency along the tumor boundary. While AU-Net++ DL2 maintains reasonable overlap,

it displays minor boundary irregularities compared to DL3. In contrast, AU-Net++ DL1 shows significant

boundary mismatches and shape distortions, indicating reduced geometric precision.

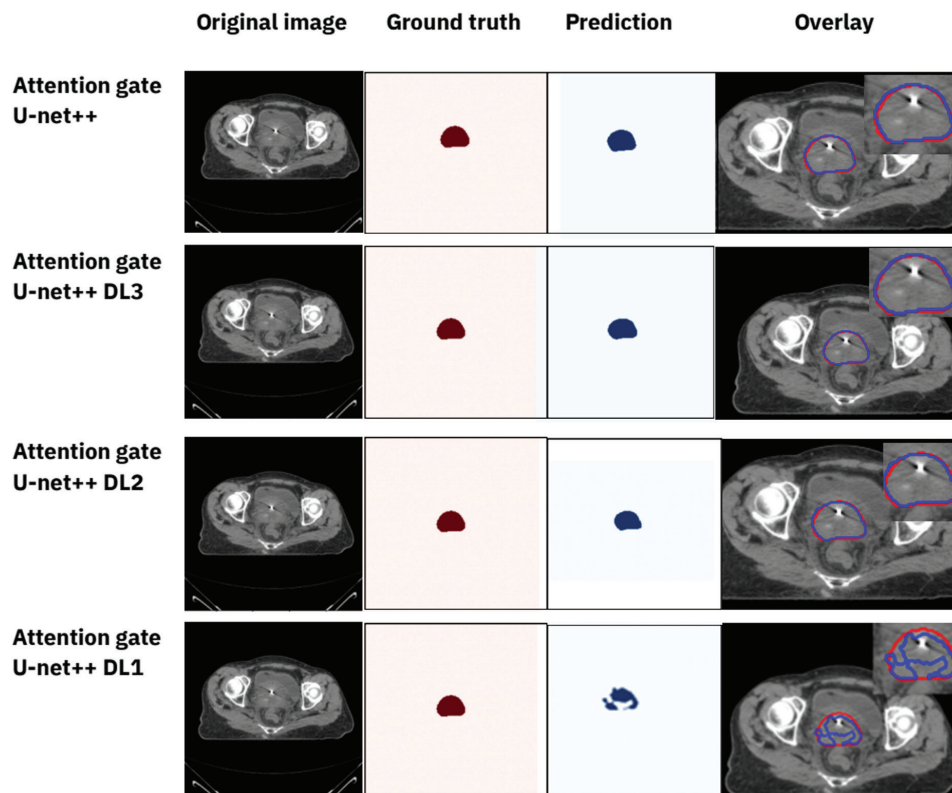


Figure 7. The boundary overlaps between the solid red contour (ground-truth) and the solid blue contour (predicted segmentation) for the Attention Gate U-Net++, AU-Net++ DL3, AU-Net++ DL2, and AU-Net++ DL1.

Discussion

This study proposes a modified Attention Gate U-Net++ architecture designed to enhance computational efficiency while maintaining segmentation performance for HR-CTV delineation in CT-based cervical cancer brachytherapy. The architectural modification entailed reducing the decoder depth (DL3) to minimize redundant feature propagation from the deepest decoder level. Tables 4 and 5 compare the proposed model with previously reported architectures, including U-Net,⁷ Attention U-Net⁸, Residual U-Net,²⁰ U-Net++,¹⁰ Residual-Attention U-Net++,¹⁶ and Attention U-Net++.¹¹ Future work should compare the proposed model with recent foundation models, such as MedSAM and other non-U-Net architectures, to further validate its generalizability and clinical applicability.

In terms of computational efficiency (Table 4), AU-Net++ DL3 reduced the number of trainable parameters from 37.30 M in the baseline Attention U-Net++ to 8.86 M, demonstrating approximately a 76% reduction. Training time decreased from 52.27 minutes to 33.11 minutes, while inference time remained comparable (36.89 and 36.97 seconds/test set). These findings indicate that decoder depth predominantly affects backpropagation complexity during training rather than forward-pass inference speed. Although DL2 (2.10 M) and DL1 (0.41 M) further reduced

model size, segmentation performance deteriorated significantly, particularly in HD95, suggesting that excessive decoder truncation compromises multi-scale feature reconstruction. Consequently, DL3 represents an optimal structural balance between model compactness and the preservation of spatial detail.

Regarding segmentation performance (Table 5), AU-Net++ DL3 achieved a DSC of 0.9076 and an IoU of 0.8376, which are comparable to Attention U-Net++ (DSC 0.9057) and conventional U-Net (0.9030). Although U-Net++ achieved the highest DSC (0.9114), the performance gain was marginal (<0.004) despite a substantially higher parameter count (36.61 M), indicating diminishing returns with increasing model complexity. Boundary accuracy (HD95=1.3959) remained close to Attention U-Net++ (1.3014) and U-Net (1.2877), suggesting that removing the deepest decoder layer did not significantly compromise contour precision. Pixel-level metrics (accuracy 0.9996, precision 0.9046, recall 0.9298) further confirm stable classification performance.

All experiments were conducted using an NVIDIA A100 GPU runtime on Google Colab, allowing all models to be trained within one hour. However, training the baseline Attention U-Net++ on a local workstation (MSI Katana, Intel Core i7 with NVIDIA RTX 4060) required approximately six hours under identical settings. This highlights the substantial influence of

hardware configuration on deep learning training time and reinforces the practical importance of architectural simplification in resource-limited environments.

Hyperparameter optimization determined that 60 epochs with a batch size of 32 provided stable convergence without overfitting. Increasing the number of epochs extended training time without meaningful improvement in DSC, supporting the study's objective of reducing computational cost while maintaining

geometric accuracy.

Overall, the deepest decoder level in Attention U-Net++ contributes limited incremental performance relative to its computational burden. The proposed AU-Net++ DL3 achieves a favorable trade-off between segmentation accuracy and efficiency, supporting its feasibility for clinical implementation, particularly in settings with constrained computational resources.

Table 4. Comparison of execution time and parameter results with previous work.

Author/year	Model	Parameters	Training time (minutes)	Inference time (seconds/test set)
Ronneberger O/ 2015 ⁷	U-net	31.37 M	16.48	36.1248
Oktay O/2018 ⁸	Attention U-net	31.90 M	19.43	35.6353
Zhang Z/2018 ²⁰	Residual U-net	33.15 M	22.80	35.7755
Zhou Z/2018 ¹⁷	U-net++	36.61 M	41.51	36.1453
Li Z/2022 ¹⁶	Residual-Attention U-Net++	39.26 M	60.90	38.1939
Yu Y/2023 ¹¹	Attention U-net++	37.30 M	52.27	36.9660
This study	Attention U-net++ DL3	8.86 M	33.11	36.8953
	Attention U-net++ DL2	2.10 M	19.41	35.2526
	Attention U-net++ DL1	0.41 M	09.26	33.2896

Note: M: millions.

Table 5. Comparison of geometric and performance results with previous work.

Author/Year	Model	DSC	IoU	HD95	Accuracy	Precision	Recall
Ronneberger O/ 2015 ⁷	U-net	0.9030	0.8296	1.2877	0.9996	0.9224	0.9041
Oktay O/2018 ⁸	Attention U-net	0.9095	0.8394	1.2481	0.9997	0.9298	0.9061
Zhang Z/2018 ²⁰	Residual U-net	0.8967	0.8190	1.3192	0.9996	0.9203	0.8932
Zhou Z/2018 ¹⁷	U-net++	0.9114	0.8424	1.4610	0.9997	0.9077	0.9313
Li Z/2022 ¹⁶	Residual-Attention U-Net++	0.9024	0.8272	2.6574	0.9996	0.9009	0.9214
Yu Y/2023 ¹¹	Attention U-net++	0.9057	0.8338	1.3014	0.9996	0.8920	0.9382
This study	Attention U-net++ DL3	0.9076	0.8376	1.3959	0.9996	0.9046	0.9298
	Attention U-net++ DL2	0.8678	0.7783	1.6236	0.9995	0.8729	0.8928
	Attention U-net++ DL1	0.7514	0.6149	3.3098	0.9991	0.9162	0.6721

Limitations

This study has several limitations that should be acknowledged. Firstly, the dataset consisted of CT images from only 56 patients collected at a single institution. Although each patient contributed multiple CT simulation sessions during brachytherapy, the relatively small sample size and single-center design may limit the generalizability of the proposed

model to other institutions with different patient populations, imaging protocols, scanner types, and clinical workflows. Therefore, external validation using larger multi-center datasets is necessary to confirm the robustness and clinical applicability of the proposed approach.

Secondly, all ground-truth segmentations

were delineated by a single experienced radiation oncologist. Although this ensured contouring consistency throughout the dataset, it did not account for inter-observer variability, which is an important consideration in HR-CTV delineation for cervical cancer brachytherapy. Future studies involving multiple radiation oncologists would enable evaluation of the model against inter-observer variation and improve its clinical reliability.

Thirdly, this study utilized only CT images for automatic HR-CTV segmentation. Although MRI examinations were performed before brachytherapy planning according to clinical guidelines, MRI-guided segmentation was not incorporated for every CT case. Consequently, the proposed model may have reduced soft-tissue discrimination compared with multimodal CT–MRI approaches. Nevertheless, CT-based brachytherapy guided by pre-treatment MRI remains a practical and clinically accepted strategy,⁵ particularly in institutions where MRI is not available for every treatment fraction.²¹ Future studies integrating multimodal CT–MRI information may further improve segmentation accuracy and anatomical boundary delineation.

Fourthly, the proposed model was developed using a two-dimensional (2D) slice-based segmentation approach. Although 2D segmentation substantially reduces computational complexity and memory requirements, it does not explicitly exploit three-dimensional spatial continuity between adjacent CT slices. As a result, volumetric consistency across slices may be compromised, and no post-processing technique was applied to enforce 3D contour continuity. Future work should investigate three-dimensional or hybrid 2D-3D architectures to better capture volumetric anatomical information while maintaining computational efficiency.

Fifthly, all CT images were acquired with a uniform slice thickness of 2 mm, which provided consistent image quality during model training but may limit generalizability to datasets acquired using different imaging resolutions. A thinner slice thickness (e.g., 1.0-1.3 mm) could improve spatial resolution and enhance geometric accuracy, particularly for boundary-sensitive metrics such as HD95. Therefore, further validation using heterogeneous imaging parameters is warranted.²²

Lastly, although the proposed decoder-reduced Attention U-Net++ substantially decreased the number of trainable parameters and reduced training time, all experiments were conducted on a high-performance NVIDIA A100 GPU. Consequently, the computational benefits of the proposed architecture have not yet been validated on lower-resource hardware commonly available in routine clinical practice. Future studies should evaluate the model on more accessible GPU platforms and perform additional external benchmarking to further demonstrate its practicality in resource-limited clinical environments.

Conclusion

This study presents a computationally efficient Attention Gate U-Net++ model for the automatic segmentation of high-risk clinical target volumes (HR-CTV) in CT-based cervical cancer brachytherapy. By reducing the decoder depth, the proposed model substantially decreases the number of trainable parameters, thereby reducing memory footprint and training cost while maintaining acceptable geometric accuracy. Despite its lightweight architecture, the proposed approach achieves competitive performance in terms of DSC, IoU, and HD95 compared with more complex models. The reduced computational demand also increases the feasibility of deploying the model on mid-range GPU platforms commonly available in clinical practice, such as the NVIDIA RTX 3060, NVIDIA RTX 4060, and NVIDIA T4, making it a practical alternative for institutions with limited computational resources. Nevertheless, additional benchmarking on these lower-resource hardware platforms is warranted to further validate its real-world computational efficiency and clinical applicability.

Ethical approval

The study was approved by the Ethics Committee of Naresuan University, Thailand (IRB No. P1-0014/2569).

Funding

None

Conflict of interest

None

CRediT authorship contribution statement

Thinnagit Srikhot: data curation, investigation, resources, writing: review and edit; **Chatrawut Pattaweerakul:** clinical validation, investigation, resources, data curation (manual contouring of HR-CTV and OAR on CT images); **Kamonchanok Nobphuek:** resources, data curation (CT image retrieval and clinical data management), validation, writing: review and edit; **Titipong Kaewlek:** conceptualization, methodology, formal analysis, writing: original draft, review and edit. All authors have read and approved the final version of the manuscript.

Acknowledgements

The authors would like to express their sincere gratitude to Assistant Professor Dr. Titipong Kaewlek for his invaluable guidance and continued support throughout this research. The authors gratefully acknowledge the medical physicist Mrs. Kamonchanok Nobphuek and radiation oncologist Dr. Chatrawut Pattaweerakul from the Department of Radiation Therapy, Naresuan University Hospital, for their kind cooperation, assistance, and facilitation in data collection.

Reference

- [1] Priyadarshini S, Swain PK, Agarwal K, Jena D, Padhee S. Trends in gynecological cancer incidence, mortality, and survival among elderly women: a SEER study. *Aging Med (Milton)*. 2024; 7(2): 179-88. doi: 10.1002/agm2.12297.
- [2] Burd EM. Human papillomavirus and cervical cancer. *Clin Microbiol Rev*. 2003; 16(1): 1-17. doi: 10.1128/CMR.16.1.1-17.2003.
- [3] Liu T, Wen S, Wang S, Yang Q, Wang X. Artificial intelligence in brachytherapy. *J Radiat Res Appl Sci*. 2024; 17(2): 100925. doi: 10.1016/j.jrras.2024.100925.
- [4] Dimopoulos JCA, Petrow P, Tanderup K, Petric P, Berger D, Kirišits C, et al. Recommendations from Gynaecological (GYN) GEC-ESTRO Working Group (IV): basic principles and parameters for MR imaging within the frame of image based adaptive cervix cancer brachytherapy. *Radiother Oncol*. 2012; 103(1): 113-22. doi: 10.1016/j.radonc.2011.12.024.
- [5] Mahantshetty U, Poetter R, Beriwal S, Grover S, Lavanya G, Rai B, et al. IBS-GEC ESTRO-ABS recommendations for CT based contouring in image guided adaptive brachytherapy for cervical cancer. *Radiother Oncol*. 2021; 160: 273-84. doi: 10.1016/j.radonc.2021.05.010.
- [6] van Vliet-Pérez SM, van Paassen R, Wauben LSGL, Straathof R, van den Berg NJ, Dankelman J, et al. Time-action and patient experience analyses of locally advanced cervical cancer brachytherapy. *Brachytherapy*. 2024; 23(3): 274-81. doi: 10.1016/j.brachy.2024.01.007.
- [7] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*. Cham: Springer; 2015: pp 234-41. doi: 10.48550/arXiv.1505.04597.
- [8] Oktay O, Schlemper J, Folgoc LL, Lee MJ, Heinrich MP, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. *arXiv*. 2018; 1804.03999. doi: 10.48550/arXiv.1804.03999.
- [9] Wang Y, Jian W, Zhu L, Cai C, Zhang B, Wang X. Attention-gated deep-learning-based automatic digitization of interstitial needles in high-dose-rate brachytherapy for cervical cancer. *Adv Radiat Oncol*. 2024; 9(1): 100941. doi: 10.1016/j.adro.2023.101340.
- [10] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer; 2018: pp 3-11. doi: 10.1007/978-3-030-00889-5_1.
- [11] Yu Y, Chen S, Wei H. Modified UNet with attention gate and dense skip connection for flow field information prediction with porous media. *Flow Meas Instrum*. 2023; 89: 102300. doi: 10.1016/j.flowmeasinst.2022.102300.
- [12] Zhong X, Ling W, Pan K, Liu C, Wu P, Zhang J, et al. A neural network architecture based on attention gate mechanism for 3D magnetotelluric forward modeling. *Comput Geosci*. 2025; 207: 106079. doi: 10.48550/arXiv.2503.11408.
- [13] Xie Z, Li X, Ma H, Wu S, Cui D. PUNet: a light-weight parallel U-Net architecture integrating Mamba-CNN for high-precision image segmentation. *Sci Rep*. 2025; 15: 38954. doi: 10.1038/s41598-025-22862-x.
- [14] Rajaramana S, Yang F, Zamzmi G, Xue Z, Antani S. Assessing the impact of image resolution on deep learning for TB lesion segmentation on frontal chest X-rays. *Diagnostics (Basel)*. 2023; 13(4): 747. doi: 10.3390/diagnostics13040747.
- [15] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019; 6: 60. doi: 10.1186/s40537-019-0197-0.
- [16] Li Z, Zhang H, Li Z, Ren Z. Residual-attention UNet++: a nested residual-attention U-Net for medical image segmentation. *Appl Sci*. 2022; 12(14): 7149. doi: 10.3390/app12147149.
- [17] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans Med Imaging*. 2019; 39(6): 1856-67. doi: 10.1109/TMI.2019.2959609.
- [18] Al Qurri A, Almekkawy M. Improved UNet with attention for medical image segmentation. *Sensors*. 2023; 23(20): 8589. doi: 10.3390/s23208589.
- [19] Zhang L, Li M, Zhang P, Liu P. EfficientSegNet: light-weight semantic segmentation with multi-scale feature fusion and boundary enhancement. *Sensors*. 2025; 25: 5934. doi: 10.3390/s25195934.
- [20] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net. *IEEE Geosci Remote Sens Lett*. 2018; 15(5): 749-53. doi: 10.1109/LGRS.2018.2802944.
- [21] Dankulchai P, Prasartseree T, Sittiwong W, Chansilpa Y, Apiwarodom N, Petsuksiri J, et al. Clinical outcome comparison between CT-guided versus all MRI-guided scenarios in brachytherapy for cervical cancer: a single-institute experience. *Clin Oncol (R Coll Radiol)*. 2024; 36(11): e456-67. doi: 10.1016/j.clon.2024.08.003.
- [22] Qing K, Yue NJ, Hathout L, Ma C, Reyhan M, Zhu J, et al. The combined use of 2D scout and 3D axial CT images to accurately determine the catheter tips for high-dose-rate brachytherapy plans. *J Appl Clin Med Phys*. 2021; 22(3): 273-8. doi: 10.1002/acm2.13184.