

Automatic delineation of organs at risk for prostate cancer on computed tomography images using convolutional neural networks

Wanitchaya Khuadpudsa¹, Sutarin Rattananon⁴, Titipong Kaewlek^{1,2,3*}

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

²Department of Radiological Technology, Faculty of Allied Health Sciences, Naresuan University, Phitsanulok, Thailand.

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

⁴Department of Radiology, Phichit Hospital, Phichit, Thailand.

ARTICLE INFO

ORCID of all authors

Wanitchaya Khuadpudsa: None

Sutarin Rattananon: <https://orcid.org/0009-0002-5821-7985>

Titipong Kaewlek: <https://orcid.org/0000-0001-6400-0489>

Correspondence

Titipong Kaewlek

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

Email: titipongk@nu.ac.th

Article history

Received 16 May 2026

Accepted as revised 8 September 2026

Available online 18 September 2026

Keywords

Auto-delineation, deep learning, segmentation, prostate cancer, CT

ABSTRACT

Background: Prostate cancer is the fourth most common cancer worldwide and the second most common in men. Accurate identification of organs at risk (OARs) is crucial for improving treatment planning, making the process faster and more efficient. Artificial intelligence has become a vital tool in medicine, reducing delays and improving treatment workflow by enabling precise, consistent automatic organ delineation.

Objectives: This study investigates and compares the performance of deep learning models for automatic OAR segmentation in CT images of prostate cancer patients.

Materials and methods: Using 2,011 CT images from the Prostate-Anatomical-Edge-Cases dataset, the U-Net, Dense V-Net, and DeepMedic models were evaluated under the same parameters for segmenting the bladder, rectum, and left and right femoral heads. Performance was assessed using the Dice Similarity Coefficient (DSC), 95th percentile Hausdorff Distance (HD95), and Mean Surface Distance (MSD).

Results: The Dense V-Net outperformed the other models, achieving the highest DSC (0.99) and the lowest HD95 (0.00) and MSD (0.00) for the bladder. U-Net performed well for well-defined structures, such as the femoral heads, while DeepMedic demonstrated the most stability.

Conclusion: Dense V-Net's superior performance highlights its effectiveness in medical image segmentation, particularly in prostate cancer treatment.

Introduction

Prostate cancer is one of the most common malignancies in men,¹ ranking second in global incidence according to the World Cancer Research Fund International.² It predominantly affects older men and continues to represent a major global health burden. Radiation therapy is a cornerstone treatment for localized and advanced prostate cancer, but its success depends on precise treatment planning that ensures maximum tumor control while minimizing damage to surrounding organs.³ In this process, accurate delineation of the prostate and organs at risk (OARs) such as the bladder, rectum, and femoral heads is essential for delivering therapeutic doses effectively while reducing radiation-induced complications that can significantly impair patients' quality of life.

Traditionally, radiation oncologists have manually delineated OARs. Although this remains the clinical standard, manual contouring is labor-intensive, time-consuming, and subject to inter- and intra-observer variability. These inconsistencies may affect treatment planning and result in inadequate tumor coverage or unnecessary irradiation of healthy tissues. Artificial intelligence (AI), particularly deep learning, has therefore emerged as a promising approach for automatic organ segmentation, improving efficiency, accuracy, and reproducibility.⁴ Recent studies have demonstrated encouraging results in deep learning-based prostate OAR segmentation; however, challenges remain because of heterogeneous tissue contrast, complex pelvic anatomy, and variations in organ shape and position. In addition, direct comparison of reported results is difficult because previous studies have used different datasets, preprocessing procedures, training strategies, and evaluation protocols.

Several deep learning architectures have been developed for medical image segmentation.⁵⁻¹⁶ U-Net is widely used because its encoder-decoder architecture

with skip connections effectively combines contextual and fine spatial information.¹⁷ However, conventional 2D U-Net processes CT images slice by slice and may have limited ability to capture inter-slice anatomical information. Dense V-Net extends the V-Net architecture by incorporating dense connections that facilitate feature propagation and mitigate vanishing gradient problems.^{18,19} DeepMedic uses a multi-scale, patch-based architecture to capture contextual information at different spatial resolutions. These models represent different approaches to feature learning and anatomical representation.^{20,21} Nevertheless, although each has demonstrated promising segmentation performance, direct comparisons of U-Net, Dense V-Net, and DeepMedic for prostate OAR delineation remain limited. It remains unclear whether their relative advantages hold under the same experimental conditions.

This gap matters because differences in datasets and experimental protocols can affect segmentation performance and obscure the network architecture's contribution. Furthermore, OARs differ in anatomical complexity: femoral heads generally have relatively distinct boundaries, whereas the rectum shows greater variation in shape, position, and surrounding tissue contrast. Therefore, overall performance alone may not adequately characterize the strengths and limitations of different architectures.

This study systematically compares U-Net, Dense V-Net, and DeepMedic for automatic segmentation of the bladder, rectum, and bilateral femoral heads on CT images for prostate cancer radiotherapy planning. Unlike previous studies that primarily evaluate individual architectures, this study compares the three models under consistent experimental conditions using the same dataset, preprocessing procedures, data partitioning, and evaluation criteria. The models are evaluated within a common 2D framework to enable

a controlled, computationally feasible comparison. Performance was evaluated using the Dice similarity coefficient (DSC), 95th percentile Hausdorff distance (HD95), and mean surface distance (MSD), providing complementary measures of segmentation overlap and boundary accuracy.

This study's original contribution is the controlled characterization of the relative performance and organ-specific strengths and limitations of these architectures, along with evaluation across different training configurations. The findings may provide practical evidence for selecting suitable deep learning architectures for efficient and reliable automated OAR delineation in prostate cancer radiotherapy planning.

Material and methods

Patient data collection

A total of 2,011 CT images of the bladder, rectum, and left and right femoral heads were obtained from the Prostate-Anatomical-Edge-Cases dataset (Cancer Imaging Archive), which comprises prostate adenocarcinoma patients treated with definitive radiotherapy between 2011 and 2019.²² Images were excluded if patients had prior bladder surgery or urinary catheters, hernias, or if images were affected by metal artifacts from surgical implants. Four organs at risk (OARs) bladder, rectum, and left and right femoral heads were selected due to their proximity to the prostate and susceptibility to radiation, making accurate delineation critical to minimize complications such as urinary toxicity, rectal injury, and femoral head necrosis. After applying the inclusion criteria, we retained 67 patient cases, yielding 2,011 CT images including bladder (600), rectum (602),

right femoral head (406), and left femoral head (403).

Deep learning models and experimental design

For this study, three representative deep learning architectures (U-Net, Dense V-Net, and DeepMedic) were selected based on their complementary strengths in medical image segmentation. To ensure a fair comparison and accommodate the 2D nature of the CT slices, all models, which were originally conceived or are often implemented in 3D, were adapted to a 2D framework.

Model architectures

Modified U-Net

The U-Net architecture, a convolutional neural network renowned for precise localization in medical imaging, was implemented with a key modification.¹⁷ Our version features a lighter encoder-decoder structure that begins with 16 filters, increases sequentially to 128 in the bottleneck, then decodes back to 16 filters. Unlike the original U-Net, which starts with 64 filters, this design significantly reduces trainable parameters and computational cost, making it well-suited to environments with limited hardware resources while preserving segmentation accuracy. The modified model in Figure 1 comprises an Encoder, Bottleneck, and Decoder. The Encoder reduces spatial dimensions while increasing filters (16→32→64→128), capturing progressively complex features. The Bottleneck learns critical representations before the Decoder restores resolution using up-sampling and skip connections. During decoding, filters are reduced (128→64→32→16) as the image size is restored to 256×256.

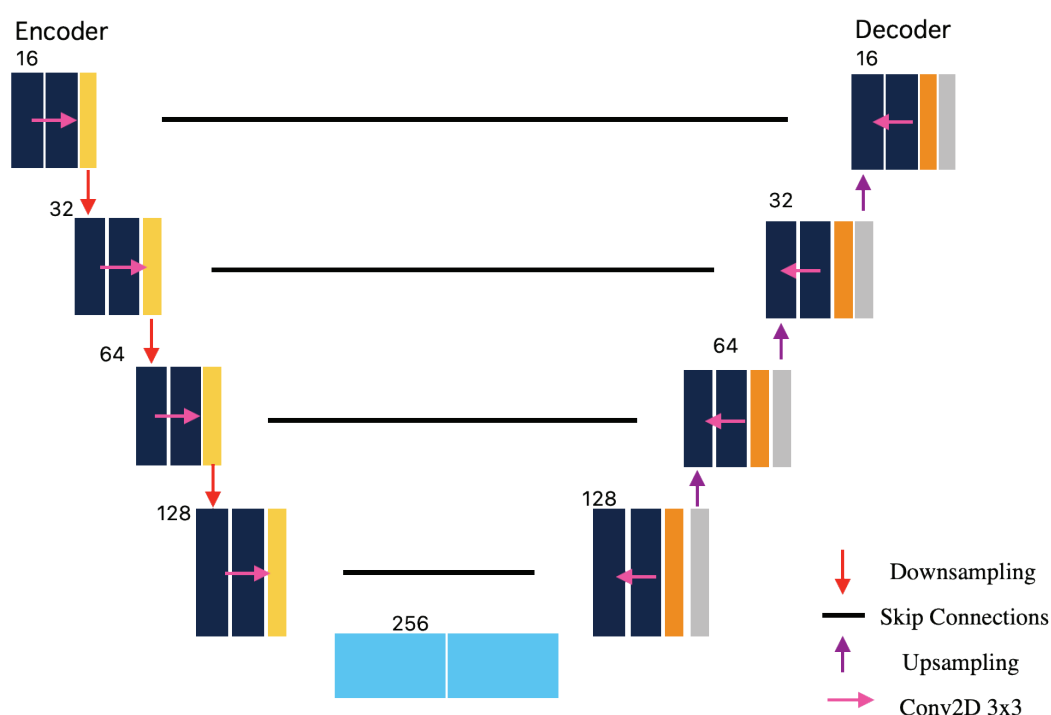


Figure 1. Architecture of the modified U-Net.

Modified Dense V-Net

We adapted the 3D Dense V-Net,²³ an extension of V-Net,³ to a 2D architecture. The model's core strength lies in its dense blocks, which facilitate strong gradient flow and extensive feature reuse through direct connections between layers. This dense connectivity is particularly effective for segmenting heterogeneous tissues. Figure 2 shows the modified Dense V-Net, adapted from the original 3D Dense V-Net by converting it to 2D for CT image segmentation.¹⁹ It consists of dense blocks, a down-sampling path, a bottleneck, and an up-sampling path. Dense blocks use stacked 2D convolutions with

ReLU activations to promote feature reuse and gradient flow. The down-sampling path extracts hierarchical features via max-pooling, while the up-sampling path restores resolution with transposed convolutions and skip connections to preserve spatial detail. Compared with the original V-Net, the modified version uses dense connectivity to improve feature propagation, reduce vanishing gradients, and improve efficiency. Compared to U-Net, Dense V-Net achieves richer feature reuse and stronger gradient flow, resulting in higher segmentation accuracy with a similar parameter count.

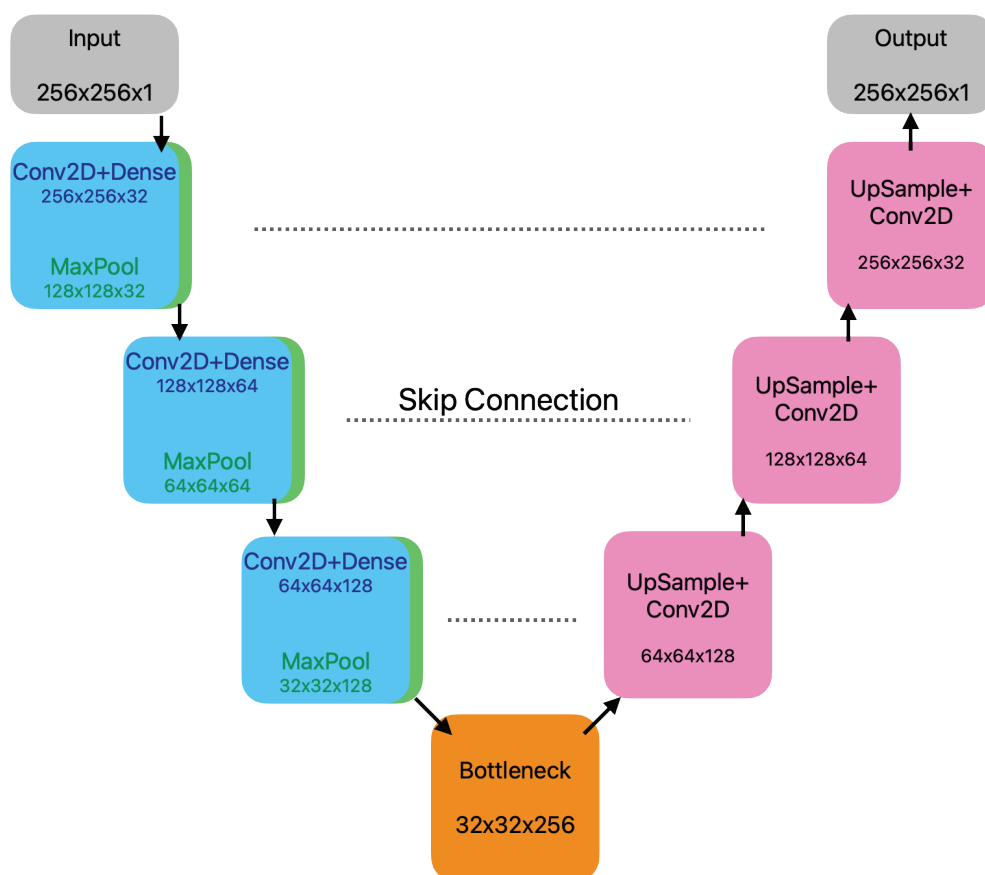


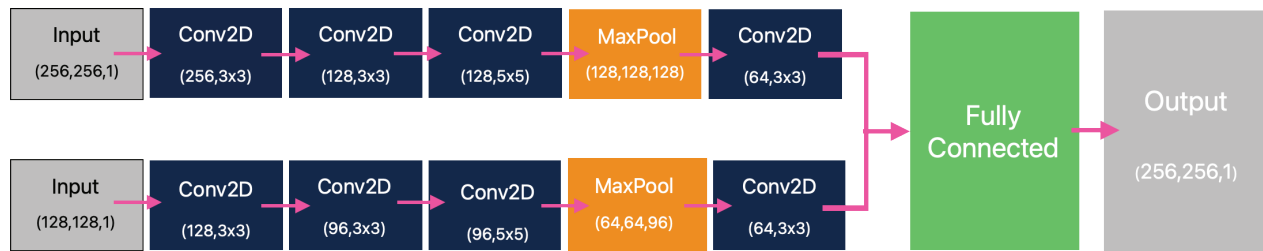
Figure 2. Architecture of the modified Dense V-Net.

Modified DeepMedic

The DeepMedic architecture, originally a 3D-CNN, was re-engineered for 2D analysis. Its defining feature is a multi-scale processing pathway.^{20,21} The model processes inputs at two resolutions simultaneously: a normal-resolution path (256×256) to capture fine details and a down-sampled, low-resolution path (128×128) to capture broader contextual information. The network merges feature from both pathways, upsamples them, and passes them through a fully connected layer to produce the final segmentation. This explicit multi-scale design distinguishes it from the implicit hierarchical feature extraction of U-Net and Dense V-Net, offering a potential advantage in

delineating structures by integrating both local and global context. Figure 3 shows that the model includes a Normal Resolution Path, a Low-Resolution Path, a Merged Path, a Fully Connected Layer, and an Output Layer. The Normal Resolution Path processes full-resolution images (256×256) with 3×3 and 5×5 convolutions, while the Low-Resolution Path captures broader context from down-sampled images (128×128). Features from both paths are merged, upsampled to 256×256, and passed through a fully connected layer before generating segmentation outputs. Compared to the original 3D DeepMedic, this 2D version reduces computational cost while retaining multi-scale feature extraction.

Normal Res Path



Low Res Path

Figure 3. Architecture of the modified model.

Preprocessing and study design

All CT images were resampled to an isotropic voxel size of $1 \times 1 \times 1 \text{ mm}^3$ and cropped or padded to 256×256 pixels. Intensities were normalized to $(-1, 1)$ without additional augmentation. A radiologist with over 15 years of experience delineated the bladder, rectum, and left and right femoral heads. The dataset was split into 70% training and 30% testing. Models were implemented on Google Colaboratory with hyperparameters tuned across batch sizes 4-15 and epochs 50-200.

Model training

Model development, training, and evaluation were conducted in Python using Google Colaboratory. Three CNN architectures, U-Net, Dense V-Net, and DeepMedic, were implemented for image segmentation. The CT slices were randomly divided at the slice level into 70% for training and 30% for testing. During model training, 20% of the training subset was reserved for validation using “validation_split = 0.2”, resulting in approximate proportions of 56% training, 14% validation, and 30% testing data. The models were compiled with the Adam optimizer, binary cross-entropy loss, and accuracy as the training metric. Training was performed using the specified combinations of batch sizes (2, 4, 8, 10, 12, 15) and epochs (50, 100, 200), and we checked the model architecture before training. Trained models were saved with identifiers reflecting their parameters. Performance was evaluated on the test set using statistical metrics, and the procedure was repeated for all three models to allow direct comparison.

Evaluation

The deep learning model for defining risk organ regions from this study was tested on the prepared test dataset. The testing process begins by loading the saved model. Load data for testing, create contours from the predicted mask, create contours from ground truth, and calculate average values used in testing,²³ including DSC, HD95, and MSD values.

Dice similarity coefficient (DSC)

DSC measures the overlap between predicted and reference segmentations and ranges from 0 to 1, with

values above 0.7 generally indicating good performance.³ Since it provides no spatial information, it should not be used alone.^{3,24} The DSC value is defined as:

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (1)$$

where X and Y are the predicted and reference masks, and $|X \cap Y|$ is their overlap. A value of 1 indicates perfect agreement.

95th percentile Hausdorff distance (HD95)

HD95 is the 95th percentile of distances from points on the predicted contour to the nearest points on the reference contour, used to evaluate boundary accuracy.³ The direct Hausdorff distance from X to Y is defined as the maximum distance from each point in X to its closest point in Y , as shown in Equation 2.

$$\text{HD95}(x,y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y)\} \quad (2)$$

Mean surface distance (MSD)

MSD is the average distance between the predicted and reference surfaces, quantifying boundary deviation.^{14,24} It is defined as:

$$\text{MSD}(A,B) = \frac{\sum_{a \in A_s} d(a, B_s) + \sum_{b \in B_s} d(b, A_s)}{|A_s| + |B_s|} \quad (3)$$

where $d(a, B_s) = \min_{b \in B_s} \|a - b\|$ is the Euclidean distance from point a on surface A_s to the nearest point on surface B_s . Here, $|A_s| + |B_s|$ denote the number of surface pixels in A and B .

Results

The performance of the three models was quantitatively evaluated using the average Dice Similarity Coefficient (DSC) for volumetric overlap, together with the average 95th percentile Hausdorff Distance (HD95) and Mean Surface Distance (MSD) for boundary accuracy.

U-Net demonstrated its highest average DSC performance for the femoral heads, with values ranging from 0.83 to 0.96 for the right femoral head and 0.87 to 0.94 for the left femoral head. For the bladder, average DSC values ranged from 0.86 to 0.94, whereas the

rectum showed values of 0.84-0.88 (Figure 4). Dense V-Net achieved the highest average DSC values overall, particularly for the bladder, with values ranging from 0.74 to 0.99 and a maximum of 0.99. For the right and left femoral heads, average DSC values ranged from 0.74 to 0.98 and 0.74 to 0.94, respectively, while the rectum ranged from 0.66 to 0.95. DeepMedic yielded average

DSC values of 0.87-0.90 for the right femoral head and 0.87-0.92 for the left femoral head. For the bladder and rectum, average DSC values ranged from 0.83 to 0.85 and 0.78 to 0.84, respectively. Overall, Dense V-Net achieved the highest average DSC values across most OARs, followed by U-Net, whereas DeepMedic generally showed lower average DSC values.

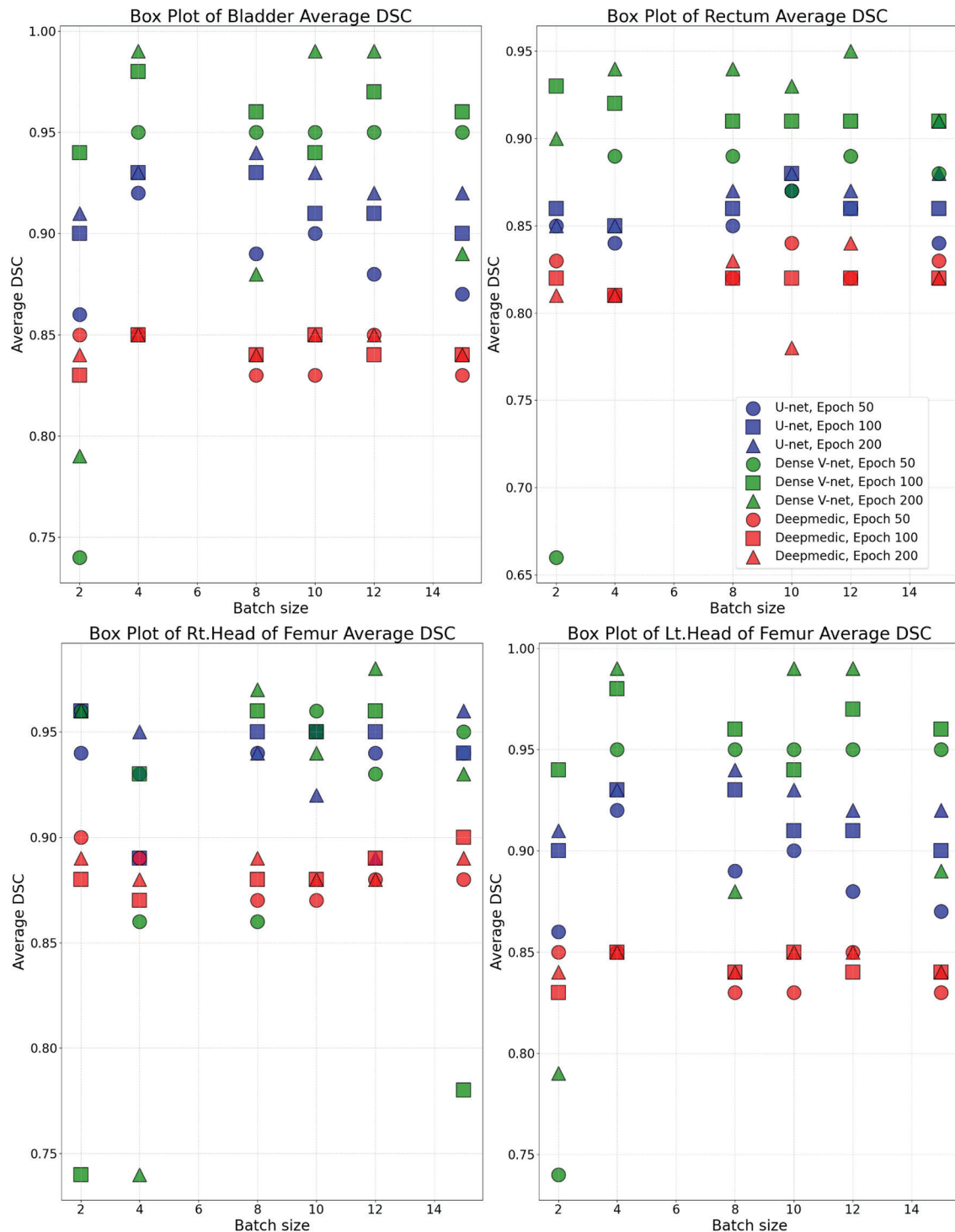


Figure 4. Box plots of average DSC values for the bladder, rectum, and femoral heads across different batch sizes and training epochs using U-Net, Dense V-Net, and DeepMedic. Marker shapes indicate the number of training epochs (50, 100, and 200), and colors represent the three network architectures.

For boundary accuracy, Dense V-Net generally achieved lower average HD95 and average MSD values (Figure 5 and Figure 6) than U-Net and DeepMedic, particularly for the bladder, where average HD95 and average MSD reached 0.00 mm (0.004749 and 0.002579) under some training configurations. However, Dense V-Net showed greater variability in boundary accuracy under specific training conditions. The highest average HD95 value observed for Dense V-Net was 8.26 mm for the right femoral head at a batch size of 2 and 100

epochs. U-Net showed comparatively consistent boundary performance across the evaluated OARs, while DeepMedic demonstrated a narrower range of DSC values and relatively consistent boundary-distance measurements. Overall, the results indicate that Dense V-Net provided the most favorable segmentation performance across the evaluated OARs, whereas U-Net showed competitive performance and DeepMedic exhibited comparatively lower volumetric overlap.

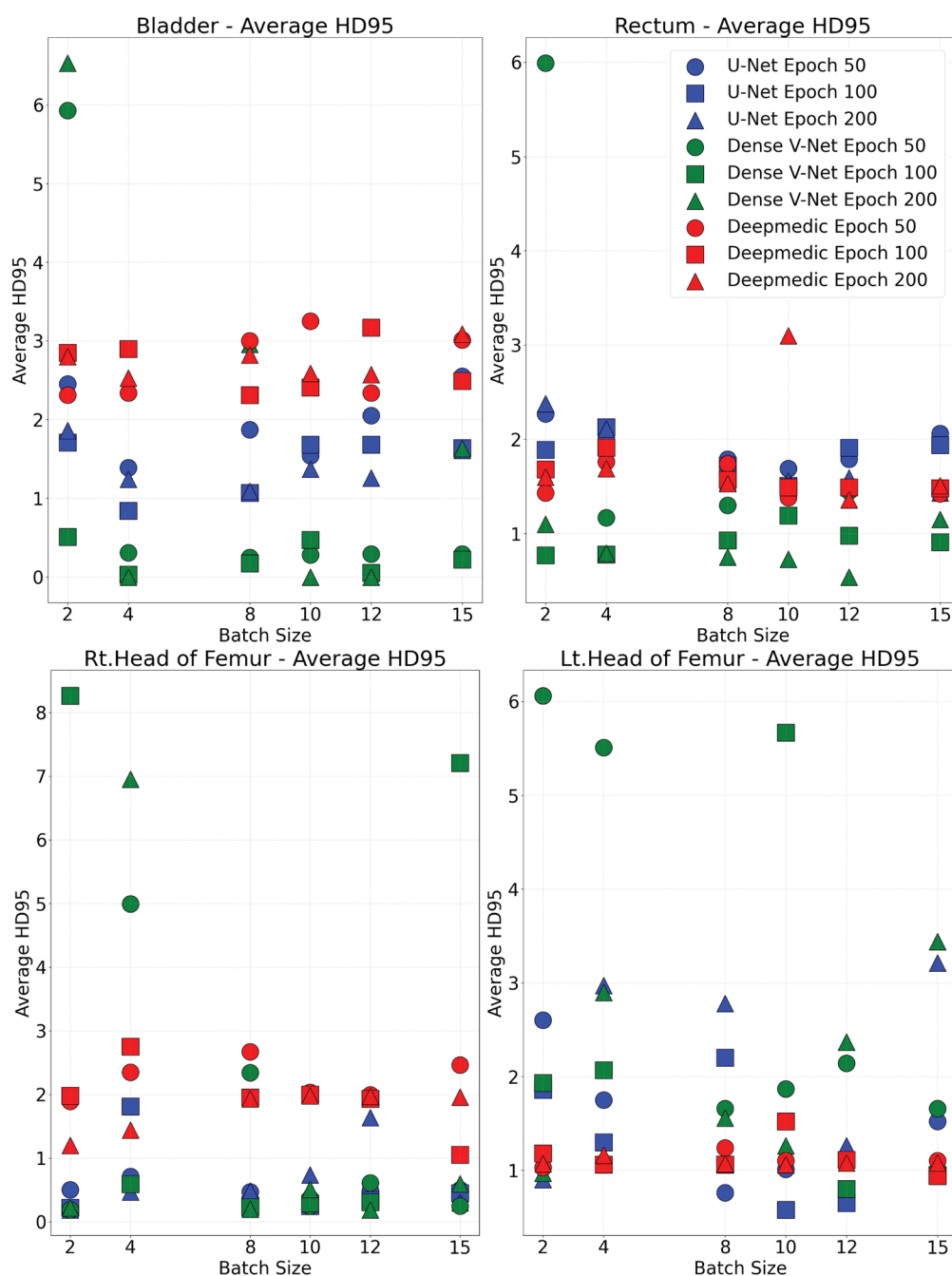


Figure 5. Box plots of average HD95 values for the bladder, rectum, and femoral heads across different batch sizes and training epochs using U-Net, Dense V-Net, and DeepMedic. Marker shapes indicate the number of training epochs (50, 100, and 200), and colors represent the three network architectures.

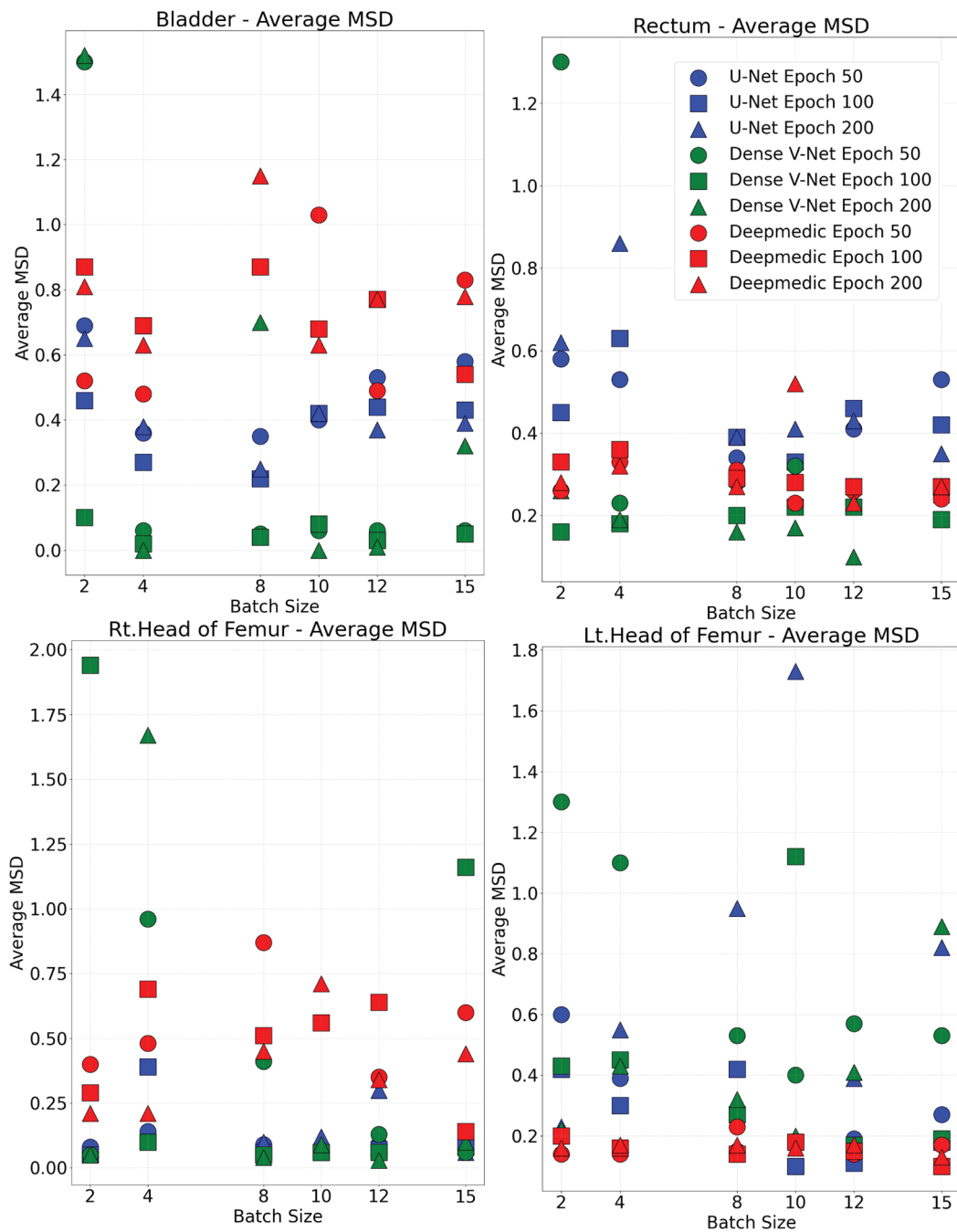


Figure 6. Box plots of average MSD values for the bladder, rectum, and femoral heads across different batch sizes and training epochs using U-Net, Dense V-Net, and DeepMedic. Marker shapes indicate the number of training epochs (50, 100, and 200), and colors represent the three network architectures.

Average HD95 and average MSD represent the distance between predicted and ground-truth segmentations; smaller values indicate better performance (Figure 5 and Figure 6). For the U-Net model (blue markers), the best-performing organs were the right and left femoral heads. The right femur achieved average HD95 values of 0.19-3.21 mm and average MSD values of 0.05-0.58 mm, while the left femur had an average HD95 of 0.25-2.55 mm and an average MSD of 0.05-1.73 mm. The bladder and rectum showed comparable results,

with an average HD95 of 0.32-2.27 mm and an average MSD of 0.32-1.42 mm. For the Deepmedic model (red markers), the left femur performed best, with an average HD95 of 0.10-1.24 mm and an average MSD of 0.10-0.94 mm. The rectum ranked second (average HD95 0.23-3.10 mm, average MSD 0.23-1.36 mm), followed by the right femur (average HD95 0.21-2.75 mm, average MSD 0.14-1.05 mm). The bladder showed the highest distances, with an average HD95 of 2.31-3.17 mm and an average MSD of 0.49-0.87 mm.

Figures 7 and Figure 8 illustrate the segmentation of the bladder, rectum, and femoral heads, comparing the ground truth (red contour) with the predicted segmentation (green contour). The top row of each figure presents the worst-performing models (lowest average DSC, highest average HD95 and average MSD), while the bottom row shows the best-performing settings. For the bladder (Figure 6), U-Net performed worst at a batch size of 2 with 50 epochs and best at a batch size of 8 with 200 epochs. Dense V-Net

performed worst at a batch size of 2 with 50 epochs and best at a batch size of 4 with 200 epochs. DeepMedic performed best at a batch size of 2 with 50 epochs. For the rectum, U-Net performed worst at a batch size of 15 with 50 epochs and best at a batch size of 15 with 200 epochs. Dense V-Net performed worst at a batch size of 2 with 50 epochs and best at a batch size of 12 with 200 epochs. DeepMedic performed best at a batch size of 12 with 200 epochs.

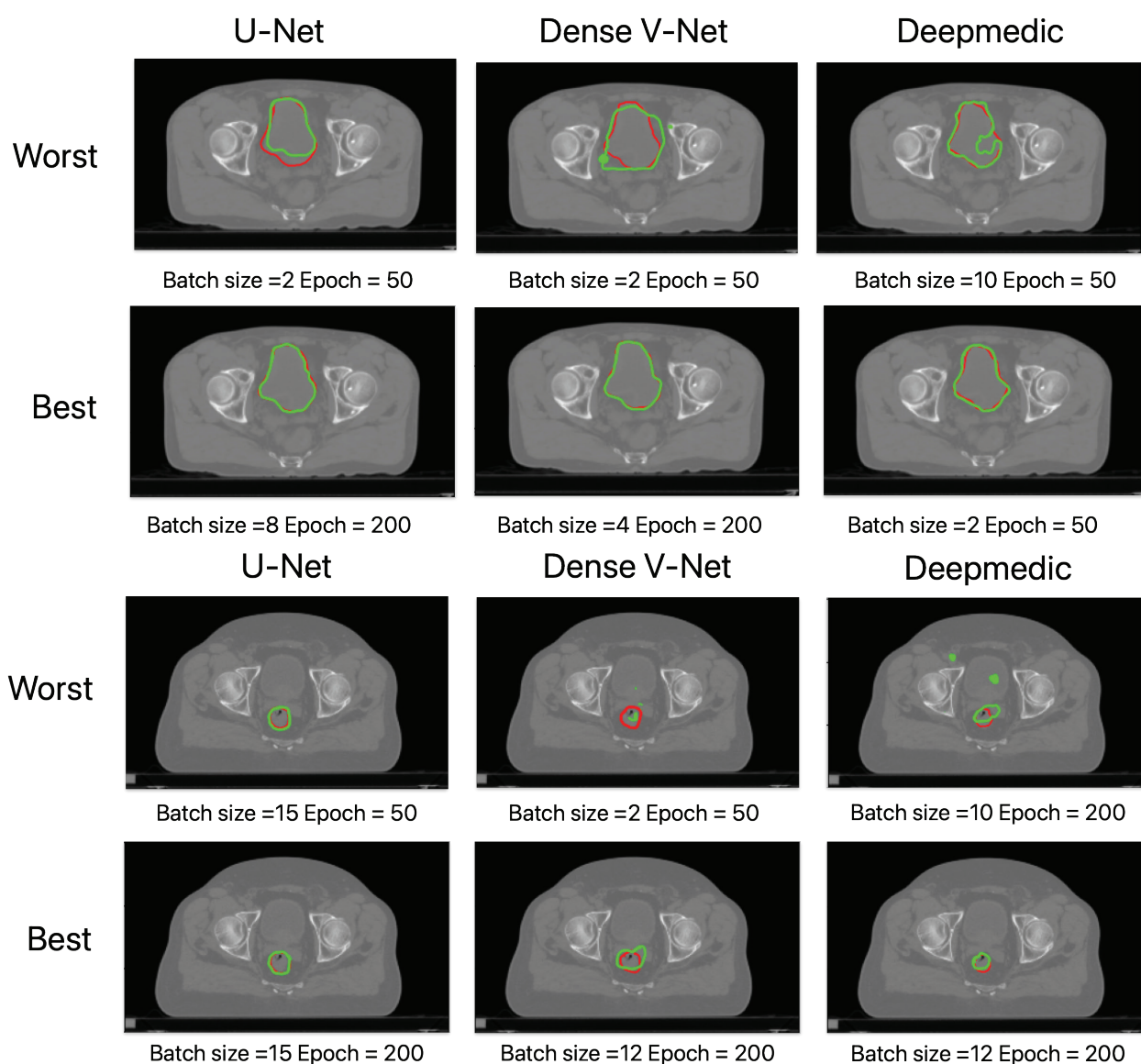


Figure 7. Image delineation of bladder and rectum. Red ground-truth contour, Green: model-predicted contour.

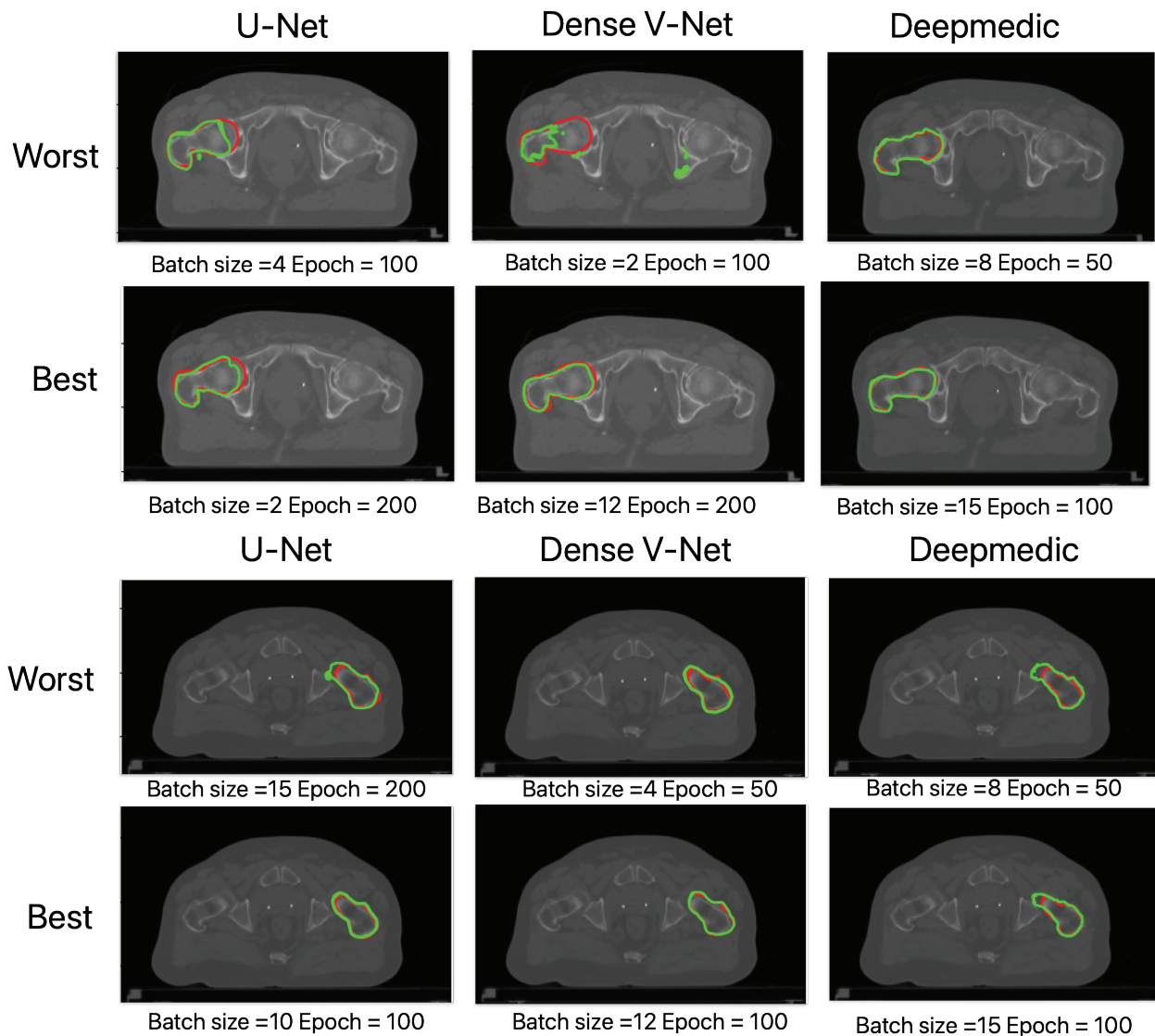


Figure 8. Image delineation of the right and left femoral heads. Red ground-truth contour, Green: model-predicted contour.

Figure 8 illustrates the segmentation evaluation for the right and left femoral heads. For the right femur, U-Net performed worst at a batch size of 4 with 100 epochs and best at a batch size of 2 with 200 epochs. Dense V-Net performed worst at a batch size of 2 with 100 epochs and best at a batch size of 12 with 200 epochs. DeepMedic performed best at a batch size of 15 with 100 epochs. For the left femur, U-Net performed worst at a batch size of 15 with 200 epochs and best at a batch size of 10 with 100 epochs. Dense V-Net performed worst at a batch size of 4 with 50 epochs and best at a batch size of 12 with 100 epochs. DeepMedic achieved its best results at a batch size of 15 with 100 epochs.

Discussion

This study evaluated three deep learning models: U-Net, Dense V-Net, and DeepMedic for automatic segmentation of organs-at-risk (OARs) in prostate cancer radiotherapy using CT images. Accurate OAR delineation is critical for treatment planning, as it

ensures precise dose delivery while sparing healthy tissue. Dense V-Net consistently outperformed the other models, achieving an average DSC of up to 0.99 and near-zero average HD95 and average MSD for the bladder, reflecting its large, well-defined anatomy. All models segmented the femoral heads accurately, given their simple, symmetrical structures. In contrast, rectum segmentation remained challenging due to its variable shape, flexible boundaries, and proximity to surrounding tissues. Parameter optimization strongly influenced segmentation accuracy. Larger batch sizes (10-12) improved performance for complex organs like the bladder and rectum, while smaller sizes (2-4) worked better for simpler structures such as the femoral heads. Increasing epochs generally improved accuracy but sometimes led to overfitting, as observed with DeepMedic, where average MSD worsened for femoral heads after prolonged training. These findings highlight the importance of careful tuning to balance learning and prevent performance degradation.

Dense V-Net's performance exceeded accepted

clinical thresholds (DSC>0.90, HD95<2-3 mm, MSD<1 mm), underscoring its clinical viability.²⁴ High-quality OAR segmentation supports patient safety by ensuring accurate dose delivery and reducing the risk of complications such as radiation-induced cystitis, proctitis, and femoral fractures.²⁵ Its robust performance is especially valuable in adaptive radiotherapy, where bladder and rectum anatomy can vary across sessions. Reliable, automated delineation enables rapid replanning, reduces clinician workload, and improves efficiency in high-volume or resource-limited settings.

U-Net offered a balance between accuracy and computational efficiency, making it suitable for faster or resource-constrained applications. DeepMedic, though less accurate in volumetric segmentation, occasionally achieved low average MSD values, indicating strong boundary detection despite weaker

overall performance. These differences reflect how model architecture, parameter tuning, and organ complexity shape segmentation outcomes.

Comparison with prior studies (Table 1) further demonstrates Dense V-Net's superiority. For the bladder and rectum, it achieved average DSC values of 0.99 and 0.95, respectively, with lower average HD95 and average MSD than previously reported using U-Net, 3D U-Net, or DeepMedic on CT,^{6,8,25-28} or MRI datasets.²⁹ For the femoral heads, average DSC values reached 0.98 (right) and 0.94 (left), outperforming earlier MRI-based studies while maintaining very low average HD95 and average MSD. These results confirm the strong performance of Dense V-Net, even in a 2D adaptation, and highlight improvements over both CT- and MRI-based approaches.

Table 1. Comparison of previous studies on automatic organ-at-risk segmentation in prostate cancer radiotherapy with average DSC, HD95, and MSD values.

Study	Modality	Model	Bladder			Rectum		
			DSC	HD95	MSD	DSC	HD95	MSD
Balogopal ²⁵	CT	U-Net	0.95	17.00	0.50	0.84	4.90	0.80
Sultana ⁸		U-Net	0.88	7.20	2.34	0.83	9.20	3.91
Nemoto ²⁶		U-Net	0.94	-	-	0.83	-	-
Wang ²⁸		U-Net	0.94	-	-	0.89	-	-
Kawula ²⁷		3D U-Net	0.96	2.50	-	0.89	5.00	-
Savenije ²⁹	MRI	Dense V-Net	0.95	3.90	1.05	0.85	7.92	2.10
		DeepMedic	0.95	3.80	0.95	0.85	8.34	2.07
This study	CT	U-Net	0.94	0.84	0.25	0.88	1.43	0.32
		Dense V-Net	0.99	0.00*	0.00**	0.95	0.54	0.10
		DeepMedic	0.85	2.31	0.49	0.84	1.36	0.23
Study	Modality	Model	Right head of femur			Left head of femur		
			DSC	HD95	MSD	DSC	HD95	MSD
Savenije ²⁹	MRI	Dense V-Net	0.97	1.50	0.50	0.97	1.60	0.50
		DeepMedic	0.96	2.15	0.61	0.96	1.89	0.55
This study	CT	U-Net	0.96	0.19	0.05	0.94	0.58	0.10
		Dense V-Net	0.98	0.19	0.03	0.94	0.80	0.17
		Deepmedic	0.90	1.05	0.14	0.92	0.94	0.10

Note: *0.004749, **0.002579

Limitations

Several limitations should be considered when interpreting the findings. Firstly, the study included 67 cases from a single dataset, which may limit the generalizability of the results across different institutions, imaging protocols, and patient populations. Secondly, the authors partitioned data at the CT-slice level, allowing images from the same patient to be distributed across different subsets. This approach may introduce data leakage and potentially result in optimistic estimates of segmentation performance. Patient-wise data partitioning, together with an independent validation dataset or nested cross-validation for model selection, would provide a more rigorous evaluation of model generalizability. Therefore, future studies should adopt patient-wise partitioning to mitigate this potential risk and provide a more robust assessment of model performance. Thirdly, Dense V-Net and DeepMedic were adapted from their original 3D architectures to a 2D framework, which may have reduced their ability to exploit inter-slice spatial information, particularly for anatomically variable structures. Evaluation of fully 3D or hybrid architectures may provide a more comprehensive assessment of volumetric contextual learning. Fourthly, we did not apply data augmentation to maintain consistent training conditions across the three architectures. Nevertheless, augmentation may improve model robustness and generalizability, particularly when patient-level sample diversity is limited. Fifthly, hyperparameter optimization was restricted to predefined combinations of batch size and training epochs. Consequently, the reported performance may not reflect the optimal performance achievable for each architecture. In addition, segmentation was analyzed using aggregate metrics rather than patient-level metrics, precluding assessment of inter-patient variability and formal statistical comparison between models. Finally, validation used data from the same dataset and did not include external validation, clinical acceptability assessment, contour-editing time analysis, or dosimetric evaluation.

Accordingly, interpret the findings as internal segmentation performance estimates rather than evidence of clinical readiness or generalizability to independent populations. Future studies should employ patient-wise data partitioning, larger multi-center datasets, external validation, systematic hyperparameter optimization, and data augmentation. Clinical evaluation should also incorporate contour-editing time, expert assessment, and dosimetric analysis. Although Dense V-Net showed the most favorable performance under the conditions evaluated, further validation is required to establish its clinical applicability.

Conclusion

Accurate segmentation of organs at risk (OARs) is critical in prostate cancer radiotherapy to ensure precise dose delivery while minimizing radiation-induced toxicity. This study compared three deep

learning models: U-Net, Dense V-Net, and DeepMedic for automatic delineation of the bladder, rectum, and femoral heads on CT images. Dense V-Net consistently performed best, with bladder average DSC up to 0.99 and average HD95 near 0 mm; rectum average DSC of 0.95 with average HD95 of 0.54 mm; and femoral head average HD95 and average MSD as low as 0.20 mm and 0.04-0.10 mm. These results exceed clinically acceptable thresholds (DSC>0.90, HD95<2 mm, MSD<1 mm), confirming Dense V-Net as the most clinically reliable model. While U-Net offered a good balance between accuracy and efficiency, and DeepMedic showed stable but lower performance, Dense V-Net outperformed both across all organs. Its consistent accuracy supports integration into clinical workflows, particularly for automated and adaptive radiotherapy planning. Future work should focus on improving model generalizability and efficiency to enhance clinical applicability.

Ethics approval

This study was approved by the Ethics Committee of Naresuan University, Thailand, Approval No. IRB No. P1-0164/2567. All procedures were performed in accordance with institutional guidelines and the Declaration of Helsinki.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Funding

None

Conflict of Interest

The authors declare no conflicts of interest or competing interests.

CRedit authorship contribution statement

Wanitchaya Khuadpudsa: conception, design, material preparation, data collection, analysis, writing-review and approval; **Sutarin Rattananon**: conception, design, material preparation, data collection, analysis, writing-original draft, review and approval; **Titipong Kaewlek**: conception, design, material preparation, data collection, analysis, writing-review and approval, supervision.

Acknowledgements

The authors thank the Department of Radiological Technology, Faculty of Allied Health Sciences, Naresuan University, for providing research facilities and technical support. The authors also thank the Medical Physicists and Radiologists who assisted with data annotation and quality assurance for this study.

References

- [1] Bott SR, Ng KL. Prostate cancer. Brisbane: Exon Publications; 2021.
- [2] World Cancer Research Fund International. Global cancer incidence in men [Internet]. 2020 [cited 2023 May 2]. Available from: www.wcrf.org/cancer-trends/worldwide-cancer-data/.
- [3] Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Kikinis R, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*. 2004; 11(2): 178-89. doi: 10.1016/S1076-6332(03)00671-8.
- [4] Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiother Oncol*. 2016; 121(2): 169-79. doi: 10.1016/j.radonc.2016.09.009.
- [5] Wong J, Huang V, Wells D, Giambattista J, Otto K, Alexander A, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiat Oncol*. 2021; 16(1): 101. doi: 10.1186/s13014-021-01831-4.
- [6] Wang J, Chen Y, Xie H, Luo L, Tang Q. Evaluation of auto-segmentation for EBRT planning structures using deep learning-based workflow on cervical cancer. *Sci Rep*. 2022; 12(1): 13650. doi: 10.1038/s41598-022-18084-0.
- [7] Tataru OS, Vartolomei MD, Rassweiler JJ, Virgil O, Lucarelli G, Ferro M, et al. Artificial intelligence and machine learning in prostate cancer patient management-current trends and future perspectives. *Diagnostics (Basel)*. 2021; 11(2): 354. doi: 10.3390/diagnostics11020354.
- [8] Sultana S, Robinson AR, Song DY, Lee J. Automatic multi-organ segmentation in computed tomography images using hierarchical convolutional neural network. *J Med Imaging (Bellingham)*. 2020; 7(5): 055001. doi: 10.1117/1.JMI.7.5.055001.
- [9] Rouvière O, Moldovan PC, Vlachomitrou A, Gouttard S, Riche B, Rabilloud M, et al. Combined model-based and deep learning-based automated 3D zonal segmentation of the prostate on T2-weighted MR images: clinical evaluation. *Eur Radiol*. 2022; 32(5): 3248-59. doi: 10.1007/s00330-021-08408-5.
- [10] Polymeri E, Johnsson ÅA, Enqvist O, Ulén J, Pettersson N, Kjölhede H, et al. Artificial intelligence-based organ delineation for radiation treatment planning of prostate cancer on computed tomography. *Adv Radiat Oncol*. 2024; 9(3): 101383. doi: 10.1016/j.adro.2023.101383.
- [11] Russell SJ, Norvig P. Artificial intelligence: a modern approach. 4th ed. Hoboken: Pearson; 2021.
- [12] Mehravand S, Yang D, Harmon SA, Xu D, Xu Z, Turkbey B, et al. Deep learning-based artificial intelligence for prostate cancer detection at biparametric MRI. *Abdom Radiol (NY)*. 2022; 47(4): 1425-34. doi: 10.1007/s00261-022-03419-2.
- [13] Isaksson LJ, Pepa M, Zaffaroni M, Marvaso G, Alterio D, Jereczek-Fossa BA, et al. Machine learning-based models for prediction of toxicity outcomes in radiotherapy. *Front Oncol*. 2020; 10: 790. doi: 10.3389/fonc.2020.00790.
- [14] Isaksson LJ, Pepa M, Summers P, Zaffaroni M, Vincini MG, Jereczek-Fossa BA, et al. Comparison of automated segmentation techniques for magnetic resonance images of the prostate. *BMC Med Imaging*. 2023; 23(1): 32. doi: 10.1186/s12880-023-00974-y.
- [15] Hirashima H, Nakamura M, Baillehache P, Fujimoto Y, Nakagawa S, Mizowaki T, et al. Development of in-house fully residual deep convolutional neural network-based segmentation software for the male pelvic CT. *Radiat Oncol*. 2021; 16(1): 135. doi: 10.1186/s13014-021-01867-6.
- [16] Elguindi S, Zelefsky MJ, Jiang J, Veeraraghavan H, Deasy JO, Hunt MA, et al. Deep learning-based auto-segmentation of targets and organs-at-risk for magnetic resonance imaging only planning of prostate radiotherapy. *Phys Imaging Radiat Oncol*. 2019; 12: 80-6. doi: 10.1016/j.phro.2019.11.006.
- [17] 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. p. 234-41.
- [18] Lempart M, Nilsson MP, Scherman J, Gustafsson CJ, Nilsson M, Olsson LE, et al. Pelvic U-Net: multi-label semantic segmentation of pelvic organs at risk for radiation therapy anal cancer patients using a deeply supervised shuffle attention convolutional neural network. *Radiat Oncol*. 2022; 17(1): 114. doi: 10.1186/s13014-022-02088-1.
- [19] Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, et al. Automatic multi-organ segmentation on abdominal CT with dense V-networks. *IEEE Trans Med Imaging*. 2018; 37(8): 1822-34. doi: 10.1109/TMI.2018.2806309.
- [20] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Glocker B, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017; 36: 61-78. doi: 10.1016/j.media.2016.10.004.
- [21] Kamnitsas K, Ferrante E, Parisot S, Ledig C, Nori AV, Antonio C, et al. DeepMedic for brain tumor segmentation. In: Crimi A, Menze B, Maier O, Reyes M, Winzeck S, Handels H, editors. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer; 2016. p. 138-49.
- [22] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013; 26(6): 1045-57. doi: 10.1007/s10278-013-9622-7.
- [23] Nai YH, Teo BW, Tan NL, O'Doherty S, Stephenson MC, Reilhac AA, et al. Comparison of metrics for

- the evaluation of medical segmentations using prostate MRI dataset. *Comput Biol Med.* 2021; 134: 104497. doi: 10.1016/S0010-4825(21)00346-6.
- [24] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging.* 2015; 15: 29. doi: 10.1186/s12880-015-0068-x.
- [25] Balagopal A, Kazemifar S, Nguyen D, Lin MH, Hannan R, Jiang S, et al. Fully automated organ segmentation in male pelvic CT images. *Phys Med Biol.* 2018; 63(24): 245015. doi: 10.1088/1361-6560/aaf11c.
- [26] Nemoto T, Futami N, Yagi M, Kunieda E, Akiba T, Shigematsu N, et al. Simple low-cost approaches to semantic segmentation in radiation therapy planning for prostate cancer using deep learning with non-contrast planning CT images. *Phys Med.* 2020; 78: 93-100. doi: 10.1016/j.ejmp.2020.09.004.
- [27] Kawula M, Purice D, Li M, Vivar G, Ahmadi SA, Kurz C, et al. Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer. *Radiat Oncol.* 2022; 17(1): 21. doi: 10.1186/s13014-022-01985-9.
- [28] Wang S, He K, Nie D, Zhou S, Gao Y, Shen D. CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. *Med Image Anal.* 2019; 54: 168-78. doi: 10.1016/j.media.2019.03.003.
- [29] Savenije MHF, Maspero M, Sikkens GG, van der Voort van Zyp JRN, Kotte ANTJ, Bol GH, et al. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. *Radiat Oncol.* 2020; 15(1): 104. doi: 10.1186/s13014-020-01528-0.