fortune Journals

Research Article

JOURNAL OF SURGERY AND RESEARCH

ISSN: 2640-1002

Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications

Tanya Fernández-Fernández[1,2], Lucía Cubero[3,4], Carmen Morote-García[3], Ana Álvarez González[2,5], Mercedes Muñóz-Fernández[5], Lydia Mediavilla-Santos[1,2,6], Rubén Pérez-Mañanes[1,2,6,7], Javier Pascau[3,7], José Antonio Calvo-Haro[1,2,6,7]

Abstract

Background: Accurate segmentation of pelvic bone tumors is crucial for effective treatment planning in radiotherapy and surgical interventions. Manual segmentation is labor-intensive and subject to variability, highlighting the need for automated solutions. This study evaluated the performance of four deep learning (DL) frameworks- U-Net, SegResNet, UNETR, and SwinUNETR-in automating the segmentation of pelvic bone tumors in a high- Complexity hospital setting, aiming to identify the most viable options for clinical integration.

Methods: A cohort of 78 patients with pelvic bone tumors from a tertiary care hospital, including patients aged 14-88 years, was used. The dataset underwent preprocessing, involving DICOM to NIfTI format conversion and focused cropping on tumor regions. These data were then divided into training, validation, and test sets. Each DL framework was trained on the same pre-processed data, with variations in hyperparameters such as image size, batch size, and data augmentation, to optimize performance. The models were evaluated based on the Dice similarity coefficient (DSC), 95% Hausdorff distance (95% HD), and average surface distance (ASD), along with training time and qualitative visual assessment.

Results: Among the four frameworks, U-Net and SwinUNETR demonstrated the best balance between segmentation accuracy and computational efficiency. U-Net achieved a DSC of (81.79 ± 21.84)% with training times of 15 minutes and 36 seconds, making it particularly suitable for environments with limited computational resources. SwinUNETR, despite longer training times, delivered the highest DSC of (82.08 ± 0.23)%. Visual evaluations confirmed that SwinUNETR and UNETR indeed provided the most visually accurate segmentations, closely aligning with the ground truth.

Conclusions: U-Net and SwinUNETR are identified as the most clinically viable DL frameworks for pelvic bone tumor segmentation, offering an optimal balance between accuracy, Computational efficiency and resource demands. Despite limitations in GPU memory and dataset size, this study contributes to the integration of automated segmentation into clinical workflows. These findings provide a strong foundation for further optimization of these models and their scalability across different tumor types, aiming to enhance patient care in oncology and improve medical imaging practices.

Keywords: Artificial intelligence; Deep Learning; UNet; SegResNet; UNETR; SwinUNETR; Automated segmentation; Pelvic Bone Sarcoma; Radiotherapy; Convolutional Neural Networks.

Affiliation:

[1]Department of Orthopaedic Surgery and Traumatology – Musculoskeletal Oncology Division. Hospital General Universitario Gregorio Marañón. Dr. Esquerdo 46, 28007.Madrid, Spain.

[2]Faculty of Medicine, Universidad Complutense. Madrid, Spain

[3]Department of Bioengineering. Universidad Carlos III de Madrid. 28911. Leganés, Spain

[4]Université Rennes, CLCC Eugène Marquis, Inserm, LTSI - UMR 1099, F-35000 Rennes, France.

[5]Department of Radiotherapeutic Oncology. Hospital General Universitario Gregorio Marañón. Dr. Esquerdo 46, 28007. Madrid, Spain.

[6]Advanced Planning and 3D Manufacturing Unit (UPAM 3D). Hospital General Universitario Gregorio Marañón. Madrid, Spain.

[7]Instituto de Investigación Sanitaria Gregorio Marañón. Dr. Esquerdo 46, 28007, Madrid, Spain.

***Corresponding author:**

Tanya Fernández-Fernández: Department of Orthopaedic Surgery and Traumatology - Musculoskeletal Oncology Division. Hospital General Universitario Gregorio Marañón. Dr. Esquerdo 46, 28007. Madrid, Spain.

Citation: Tanya Fernández-Fernández, Lucía Cubero, Carmen Morote-García, Ana Álvarez González, Mercedes Muñóz-Fernández,Lydia Mediavilla-Santos, Rubén Pérez-Mañanes, Javier Pascau, José Antonio Calvo-Haro. Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications. Journal of Surgery and Research. 8 (2025): 274-284.

Received: April 07, 2025
Accepted: April 15, 2025
Published: May 26, 2025

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

Introduction

Pelvic bone tumors, encompassing a diverse range of benign and malignant neoplasms, pose significant challenges in musculoskeletal oncology because of their complex anatomical location and proximity to critical structures [1,2]. Precise identification and delineation of tumor boundaries are paramount for effective treatment planning, particularly in radiotherapy and surgical interventions [3]. Traditionally, the segmentation of these tumors has relied heavily on manual techniques, wherein radiologists and surgeons painstakingly outline tumor margins by imaging studies such as computed tomography (CT) [4] and magnetic resonance imaging (MRI)[5]. However, this manual process is not only time-consuming but also subject to significant interobserver variability, which can introduce inconsistencies in treatment planning and potentially impact clinical outcomes [6]. The rapid advancement of computer technology, particularly in the areas of computer vision, image processing, and pattern recognition, has facilitated the development of digital image segmentation as a crucial tool in medical imaging [7]. Digital image segmentation involves partitioning an image into distinct regions based on specific attributes such as color, texture, and density [8]. This enables the quantitative analysis of medical images-a critical component for accurate diagnosis, treatment planning, and disease monitoring. Deep learning (DL), a sophisticated branch of artificial intelligence (AI), has demonstrated remarkable potential across various domains, including medical imaging [9]. Convolutional neural networks (CNNs), a subset of DL algorithms, have revolutionized image segmentation by automating the process with high accuracy [10]. These algorithms are capable of learning intricate patterns and features from extensive datasets, allowing them to perform tasks that traditionally require human expertise. The application of DL to medical image segmentation has proven particularly transformative, significantly reducing the time and effort required to delineate tumor boundaries manually [11]. In radiotherapy, accurate image segmentation is indispensable [4]. Segmentation serves as a pivotal step in the radiotherapy workflow by precisely identifying the target treatment area while sparing adjacent healthy tissues from unnecessary irradiation [3]. However, the labor-intensive nature and inherent variability of manual segmentation can delay treatment initiation and adversely affect patient outcomes [6]. DL-based segmentation tools hold the promise of automating this process, thereby enhancing consistency, minimizing human error, and improving overall treatment efficacy [12,13]. These tools are assessed not only for their accuracy and efficiency in radiotherapy planning but also for their applicability in the surgical field, including 3D preoperative training, the creation of 3D-printed surgical guides, and the facilitation of surgical navigation systems. In summary, despite technological advancements in imaging, traditional manual segmentation methods remain laborious and subjective, heavily reliant on the clinician's expertise and interpretation. These tasks are often repetitive and mechanical yet crucial, making them prime candidates for automation [12,13]. To address these limitations [14,15], this study focuses on evaluating existing CNN-based automatic segmentation9 methods designed to accurately identify and segment tumors in the pelvic region, specifically pelvic bone sarcomas. By utilizing AI tools, this study aims to streamline the segmentation process, reducing the burden on clinicians and enhancing consistency in treatment planning.

Materials and Methods

Aim and specific objectives

The primary objectives of this study are as follows:

- To prepare a dataset of tumor regions segmented by specialists; anonymization and adaptation for CNN training are ensured.
- The prepared dataset is used to train and optimize four existing CNN-based segmentation frameworks.
- To evaluate the performance of each framework, the Dice score coefficient

(DSC), 95% Hausdorff distance (95% HD), and average surface distance (ASD) are used to select the framework that yields the best overall performance for clinical application. Through this comparative evaluation, the study aims to identify the most effective tool for improving the accuracy and efficiency of pelvic tumor segmentation, ultimately enhancing radiotherapy and surgical planning.

Study design and data acquisition

This study was conducted with datasets obtained from the medical database of a Musculoskeletal Sarcoma reference unit at a tertiary care hospital. Data were specifically extracted for 104 patients with pelvic bone sarcomas who received preoperative radiotherapy between 2015 and 2023. Of these, only 78 cases were deemed valid, where the gross tumor volume (GTV) had been manually segmented by radiotherapists. The dataset included both male and female patients, with 43.27% being women and 56.73% being men. The age range for females was 14-87 years, with an average age of 60.71 years, whereas for males, the age range was 14-88 years, with an average age of 64.19 years. The data, which were originally in Digital Imaging and Communications in Medicine (DICOM) format, were anonymized and converted to Neuroimaging Informatics Technology Initiative (NIfTI) format for subsequent processing and analysis[16,17]. The entire study utilized the MONAI (Medical Open Network for AI) framework, a specialized platform for DL in medical imaging, encompassing data preprocessing, model training, and evaluation [18,19].

Citation: Tanya Fernández-Fernández, Lucía Cubero, Carmen Morote-García, Ana Álvarez González, Mercedes Muñóz-Fernández, Lydia Mediavilla-Santos, Rubén Pérez-Mañanes, Javier Pascau, José Antonio Calvo-Haro. Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications. Journal of Surgery and Research. 8 (2025): 274-284.

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

Preprocessing

The preprocessing phase was critical to prepare the dataset for DL applications. Initially, DICOM files, which serve as the standard format for storing and transmitting medical images, were manually converted to NIfTI format via 3D Slicer, a comprehensive software platform for the analysis and visualization of medical images [20]. This conversion was necessary to simplify the files for deep learning while retaining the essential image data and tumor masks [16]. The preprocessing steps followed a structured pipeline to ensure data consistency and quality (Figure 1).

1. DICOM to NIfTI Conversion[16]: The CT scans, which were originally in DICOM format, were converted to the NIfTI format. Each scan had a resolution of [512×512] voxels in the x- and y-axes, with the number of slices varying depending on the tumor size, ranging from 75-389 slices. The tumor volumes ranged from 17 to 101 slices. The heterogeneity in tumor size influences the cropping technique used.

2. Cropping: Cropping involves reducing the number of slices along the z-axis to focus on the tumor region (slice cropping) and adjusting the area of each slice along the x- and y-axes to eliminate nonrelevant background information (image cropping).

3. Dataset Division: The dataset was divided into training, validation, and test sets, following an (80-10-10) % split to ensure sufficient data for training while keeping enough images for validation and testing [21].

4. **Data Augmentation[15]:** To increase model robustness, data augmentation techniques such as rotations, translations, and the addition of Gaussian noise were applied exclusively to the training dataset. The validation and test sets remained unaltered to ensure that the evaluation results were realistic and reflective of the model's performance on new data.

5. Parameter Optimization: This includes fine-tuning hyperparameters such as the learning rate, loss function, and optimizer settings to ensure optimal model performance [22]. These parameters were kept constant across all training sessions for consistency.

Figure 1: Preprocessing pipeline.

Model training and evaluation

Four CNN-based segmentation frameworks were selected for evaluation: U-Net, SegResNet, UNETR, and SwinUNETR. The models were trained on the prepared dataset, with experiments designed to optimize hyperparameters and assess the impact of preprocessing techniques on segmentation accuracy. All the models were trained on two NVIDIA TITAN X GPUs with 12 GB of RAM.Owing to memory limitations, only U-Net supported image sizes greater than [96×96×96]. The training process was managed remotely via the Secure Shell Protocol (SSH) using the Terminal application on Mac, and the code was implemented via Jupyter Notebooks. The first framework, U-Net, employs a U-shaped architecture comprising a contracting path for feature extraction and an expansive path for generating segmentation masks [10]. This design allows for the precise delineation of tumor boundaries, even with limited training data, by leveraging skip connections that combine features from different layers. Building on U-Net, SegResNet incorporates ResNet blocks within the encoder, enhancing feature extraction and providing robust performance in 3D image segmentation by mitigating the vanishing gradient problem through deep residual connections [23]. To address the limitations of traditional CNNs in capturing the global context and long-range dependencies, the UNETR framework integrates transformers into the U-Net architecture [24]. This innovation allows UNETR to effectively learn global features, making it particularly well suited for complex 3D medical image segmentation tasks. Finally, SwinUNETR represents an advanced variation of UNETR, incorporating a Swin transformer backbone. This model hierarchically processes image patches, utilizing shifted windows to capture both local and global contextual representations efficiently [25]. Initially developed for brain MRI segmentation, SwinUNETR's application to pelvic bone sarcoma segmentation was explored in this study to assess its potential in a different clinical context.

Post-processing

After training, the models produced labelled predictions where each voxel was classified as either background or tumor. A postprocessing step using MONAI's AsDiscrete transform was implemented to refine the segmentation maps by assigning each voxel to the most likely class [18]. This

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

step was essential to ensure that the segmentation outputs were accurate and ready for clinical use.

Evaluation criteria

The networks were evaluated on three key criteria: training time, geometric evaluation metrics, and visual evaluation of the tumor segmentation predicted by the model. Geometric metrics, such as the DSC, 95% HD and ASD, provide a quantitative assessment of segmentation accuracy [26]. Additionally, the visual evaluation involved a detailed inspection of the segmentation maps generated by each network to assess the alignment with actual tumor boundaries. By analysing these factors, the performance and effectiveness of the networks were thoroughly assessed, offering valuable insights into their capabilities and clinical applicability.

Results

Training Times

The training durations for each CNN architecture, specifically within the context of pelvic bone sarcomas, are detailed in figures 3-6. The experiments that achieved the highest DSCs on the test set are highlighted in bold, underscoring their superior performance. Importantly, comparisons between experiments 5 and 6 for U-Net with other architectures are limited because of the differences in image size and GPU memory constraints. In particular, the U-Net experiments involved doubling the image size, which was not feasible for the other architectures due to memory limitations. Additionally, the UNETR and SwinUNETR architectures require more epochs to reach convergence, leading to longer training times. These differences underscore the challenges of directly comparing training durations across architectures but highlight the significance of computational efficiency in clinical applications, where the speed of model training can be a critical factor.

Model evaluation

The evaluation of the CNN architectures was based on three key geometric metrics: the Dice score (DSC), 95% HD, and ASD. These metrics were computed specifically for tumor segmentation and provide a quantitative assessment of the model's performance.

U-Net

U-Net exhibited variable performance across the experiments. The DSCs for the test set ranged from 59.08% to 81.78% (Table 1). Notably, experiments that maintain a balance between the image size and the inclusion of background in the Dice metric yielded the best results. Despite its good performance, U-Net struggled with the complexity of pelvic bone sarcoma segmentation, particularly when dealing with heterogeneous image sizes and the application of cropping techniques.

Table 1: Evaluation Metrics for the U-Net Framework in Pelvic Bone Sarcoma Segmentation.

		Evaluation Metrics		
N	**Metric**	**Training**	**Validation**	**Test**
1	Dice (%)	67.157 ± 20.870	72.721 ± 22.649	61.615 ± 14.480
	Hausdorff Dist. 95 (mm)	5.566 ± 8.997	10.236 ± 11.202	8.444 ± 9.587
	Average Surf. Dist. (mm)	2.365 ± 4.127	0.731 ± 0.814	3.711 ± 4.450
2	Dice (%)	72.680 ± 21.435	72.581 ± 23.041	64.955 ± 17.689
	Hausdorff Dist. 95 (mm)	3.324 ± 6.525	9.235 ± 11.626	5.257 ± 6.103
	Average Surf. Dist. (mm)	1.022 ± 1.972	0.970 ± 1.471	2.595 ± 2.968
3	Dice (%)	96.734 ± 5.979	75.393 ± 22.442	81.500 ± 21.440
	Hausdorff Dist. 95 (mm)	0.799 ± 1.194	5.012 ± 5.012	5.329 ± 6.807
	Average Surf. Dist. (mm)	0.379 ± 1.142	1.665 ± 2.122	1.972 ± 3.150
4	Dice (%)	98.810 ± 0.481	76.000 ± 0.225	**81.788 ± 21.837**
	Hausdorff Dist. 95 (mm)	0.534 ± 0.200	8.392 ± 8.075	6.055 ± 7.2264
	Average Surf. Dist. (mm)	0.070 ± 0.035	2.963 ± 3.758	2.295 ± 3.473
5	Dice (%)	63.604 ± 10.533	58.400 ± 6.879	59.077 ± 12.075
	Hausdorff Dist. 95 (mm)	31.557 ± 14.063	33.203 ± 9.344	37.156 ± 18.142
	Average Surf. Dist. (mm)	13.309 ± 6.665	11.053 ± 2.983	14.978 ± 9.833
6	Dice (%)	87.256 ± 9.820	77.058 ± 17.495	74.895 ± 19.462
	Hausdorff Dist. 95 (mm)	19.789 ± 16.153	20.683 ± 20.681	36.047 ± 7.611
	Average Surf. Dist. (mm)	6.540 ± 8.295	4.843 ± 5.425	12.257 ± 7.757

Citation: Tanya Fernández-Fernández, Lucía Cubero, Carmen Morote-García, Ana Álvarez González, Mercedes Muñóz-Fernández, Lydia Mediavilla-Santos, Rubén Pérez-Mañanes, Javier Pascau, José Antonio Calvo-Haro. Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications. Journal of Surgery and Research. 8 (2025): 274-284.

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

SegResNet: SegResNet demonstrated robust performance, with Dice scores reaching up to 81.24% on the test set (Table 2). The architecture's incorporation of ResNet blocks enhanced its ability to manage 3D image segmentation effectively. However, some experiments revealed a decline in accuracy, potentially due to the challenging nature of the dataset and the specific training conditions applied.

Table 2: Evaluation Metrics for the SegResNet Framework in Pelvic Bone Sarcoma Segmentation.

		Evaluation Metrics		
N	**Metric**	**Training**	**Validation**	**Test**
1	Dice (%)	57.075 ± 11.036	59.290 ± 9.719	55.036 ± 12.009
	Hausdorff Dist. 95 (mm)	23.793 ± 10.096	21.958 ± 8.006	29.380 ± 10.414
	Average Surf. Dist. (mm)	11.153 ± 6.943	7.379 ± 4.718	11.882 ± 4.795
2	Dice (%)	94.984 ± 2.843	82.173 ± 13.872	80.325 ± 0.201
	Hausdorff Dist. 95 (mm)	1.007 ± 1.409	3.974 ± 3.566	3.843 ± 4.458
	Average Surf. Dist. (mm)	0.169 ± 0.057	0.827 ± 1.069	0.705 ± 0.741
3	Dice (%)	95.923 ± 1.567	66.692 ± 32.627	61.764 ± 41.222
	Hausdorff Dist. 95 (mm)	1.322 ± 1.633	7.070 ± 6.893	8.583 ± 9.969
	Average Surf. Dist. (mm)	0.135 ± 0.036	2.059 ± 3.172	3.529 ± 4.359
4	Dice (%)	97.952 ± 0.730	83.224 ± 16.415	**81.241 ± 21.094**
	Hausdorff Dist. 95 (mm)	0.661 ± 0.773	3.855 ± 3.876	4.562 ± 5.366
	Average Surf. Dist. (mm)	0.070 ± 0.022	1.058 ± 1.624	1.797 ± 2.606
5	Dice (%)	94.602 ± 6.083	86.088 ± 0.121	79.562 ± 21.734
	Hausdorff Dist. 95 (mm)	1.389 ± 1.441	4.789 ± 4.040	8.197 ± 9.479
	Average Surf. Dist. (mm)	0.199 ± 0.039	1.189 ± 1.545	3.213 ± 4.836

UNETR: The integration of transformers within the UNETR architecture allows for improved capture of the global context and long-range dependencies in images, which is particularly advantageous for complex 3D medical image segmentation tasks. UNETR achieved DSCs of up to 81.66% (Table 3), reflecting its ability to handle the intricate anatomy of pelvic bone sarcomas. Nonetheless, the higher computational demands and longer training times are important considerations for practical implementation.

Table 3: Evaluation Metrics for the UNETR Framework in Pelvic Bone Sarcoma Segmentation.

		Evaluation Metrics		
N	**Metric**	**Training**	**Validation**	**Test**
1	Dice (%)	79.866 ± 10.810	70.415 ± 13.149	73.589 ± 14.005
	Hausdorff Dist. 95 (mm)	4.613 ± 3.471	5.574 ± 3.047	5.115 ± 3.447
	Average Surf. Dist. (mm)	1.426 ± 2.301	1.534 ± 0.973	1.655 ± 1.578
2	Dice (%)	87.572 ± 9.836	74.726 ± 15.721	77.284 ± 15.835
	Hausdorff Dist. 95 (mm)	2.757 ± 3.476	4.402 ± 3.187	3.379 ± 2.784
	Average Surf. Dist. (mm)	0.944 ± 2.146	1.866 ± 2.023	1.269 ± 1.345
3	Dice (%)	95.551 ± 7.096	75.816 ± 21.509	**81.664 ± 21.216**
	Hausdorff Dist. 95 (mm)	0.514 ± 0.123	4.609 ± 4.404	4.192 ± 5.568
	Average Surf. Dist. (mm)	0.080 ± 0.029	1.929 ± 2.300	1.599 ± 3.064
4	Dice (%)	82.901 ± 8.163	71.914 ± 12.563	74.622 ± 15.385
	Hausdorff Dist. 95 (mm)	3.352 ± 2.313	6.982 ± 3.750	5.176 ± 4.509
	Average Surf. Dist. (mm)	0.748 ± 0.426	2.178 ± 1.803	1.901 ± 2.908
5	Dice (%)	90.988 ± 8.187	76.442 ± 16.032	77.950 ± 21.732
	Hausdorff Dist. 95 (mm)	1.624 ± 0.699	8.217 ± 7.264	7.857 ± 8.671
	Average Surf. Dist. (mm)	0.363 ± 0.079	1.813 ± 2.039	2.923 ± 4.032
6	Dice (%)	60.714 ± 6.233	66.669 ± 13.382	62.481 ± 12.865
	Hausdorff Dist. 95 (mm)	21.073 ± 4.059	15.356 ± 4.849	17.104 ± 4.726
	Average Surf. Dist. (mm)	7.727 ± 2.170	4.882 ± 2.504	5.594 ± 2.487

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

SwinUNETR: SwinUNETR, an advanced variation of UNETR that incorporates a Swin transformer backbone, showed promising results, with Dice scores reaching up to 82.08% on the test set (Table 4). This model's ability to efficiently process image patches hierarchically and capture both local and global contextual representations made it particularly well suited for the segmentation of pelvic bone sarcomas. However, similar to UNETR, SwinUNETR's increased computational complexity and training times must be considered in its potential clinical application.

Table 4: Evaluation Metrics for the SwinUNETR Framework in Pelvic Bone Sarcoma Segmentation.

		Evaluation Metrics		
N	**Metric**	**Training**	**Validation**	**Test**
1	Dice (%)	93.195 ± 6.880	74.053 ± 19.501	80.434 ± 18.616
	Hausdorff Dist. 95 (mm)	1.461 ± 2.845	5.343 ± 4.947	4.758 ± 5.510
	Average Surf. Dist. (mm)	0.337 ± 0.581	2.222 ± 2.196	1.647 ± 2.566
2	Dice (%)	97.915 ± 1.023	75.526 ± 20.505	81.025 ± 21.502
	Hausdorff Dist. 95 (mm)	0.518 ± 0.215	5.458 ± 5.151	5.031 ± 6.532
	Average Surf. Dist. (mm)	0.085 ± 0.049	1.896 ± 1.891	2.188 ± 3.427
3	Dice (%)	99.408 ± 0.431	78.070 ± 21.291	**82.080 ± 0.225**
	Hausdorff Dist. 95 (mm)	0.273 ± 22.060	4.669 ± 4.545	5.235 ± 7.412
	Average Surf. Dist. (mm)	0.024 ± 0.017	1.840 ± 1.893	2.367 ± 3.840
4	Dice (%)	94.764 ± 6.879	77.109 ± 19.380	80.741 ± 20.880
	Hausdorff Dist. 95 (mm)	0.687 ± 0.812	5.037 ± 4.925	5.295 ± 7.185
	Average Surf. Dist. (mm)	0.112 ± 0.028	1.874 ± 1.912	2.072 ± 3.293
5	Dice (%)	94.525 ± 5.859	75.988 ± 19.366	80.398 ± 20.416
	Hausdorff Dist. 95 (mm)	1.134 ± 1.425	8.387 ± 8.001	6.403 ± 7.597
	Average Surf. Dist. (mm)	0.191 ± 0.036	2.929 ± 3.154	2.505 ± 3.601
6	Dice (%)	73.788 ± 15.866	74.529 ± 13.245	73.788 ± 15.866
	Hausdorff Dist. 95 (mm)	8.547 ± 4.577	10.292 ± 8.084	8.547 ± 4.577
	Average Surf. Dist. (mm)	2.582 ± 2.243	2.085 ± 1.659	2.582 ± 2.243

Visual evaluation

In addition to geometric metrics, visual evaluation was conducted to assess the quality of the predicted segmentation. figures (2-5) present the ground truth and predicted segmentations for a selected pelvic bone sarcoma case. The visual comparison demonstrates the models' abilities to capture the overall shape and structure of the tumor, with UNETR (Figure 8) and SwinUNETR (Figure 2) providing the most visually accurate segmentations, closely aligning with the ground truth.

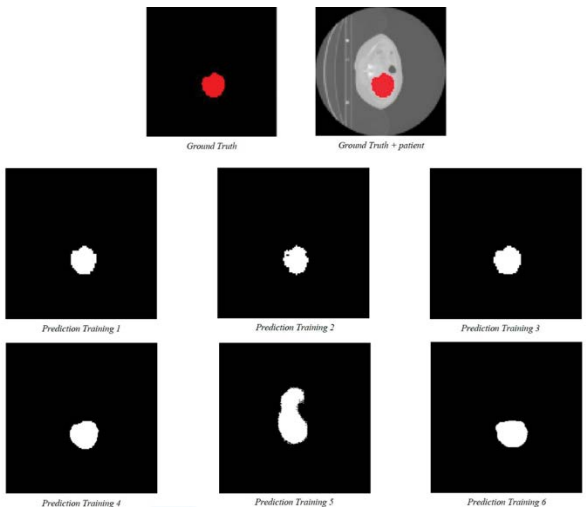

Figure 2: U-Net model prediction vs. ground truth for pelvic bone sarcoma segmentation.

Citation: Tanya Fernández-Fernández, Lucía Cubero, Carmen Morote-García, Ana Álvarez González, Mercedes Muñóz-Fernández, Lydia Mediavilla-Santos, Rubén Pérez-Mañanes, Javier Pascau, José Antonio Calvo-Haro. Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications. Journal of Surgery and Research. 8 (2025): 274-284.

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

This image compares the ground truth segmentation of a pelvic bone sarcoma (top row) with the segmentation predictions generated by the U-Net model across six different training scenarios (bottom two rows). The ground truth image shows the actual tumor segmentation, whereas the adjacent image overlays the tumor on the patient's CT scan for reference. The six prediction images demonstrate how the model's accuracy varies depending on the specific training conditions, highlighting differences in tumor shape and boundary delineation.

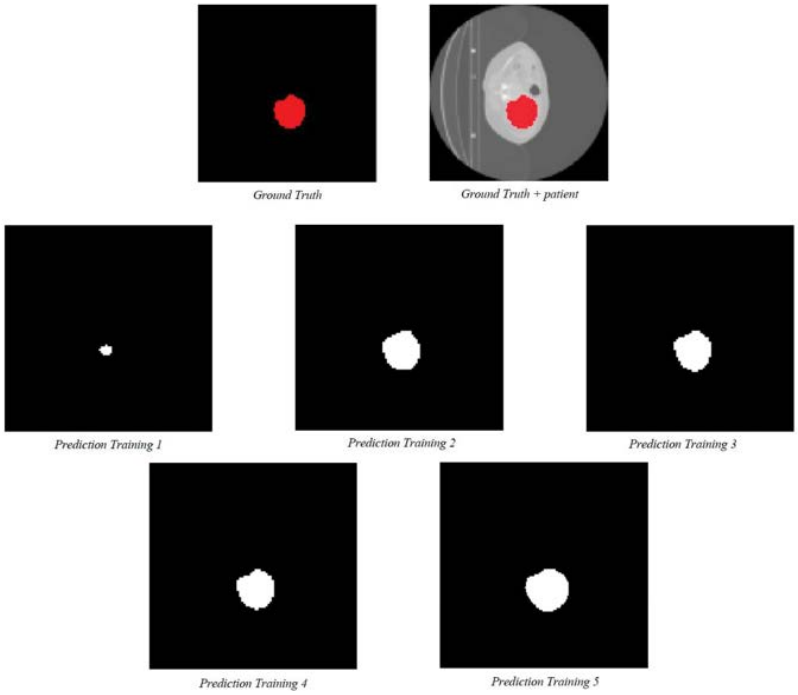

Figure 3: Visual representation of the SegResNet results.

This image shows the ground truth segmentation of a pelvic bone sarcoma (top row) compared with the SegResNet model predictions from six training scenarios (bottom two rows). The ground truth and its overlay on the patient's CT scan are provided for reference.

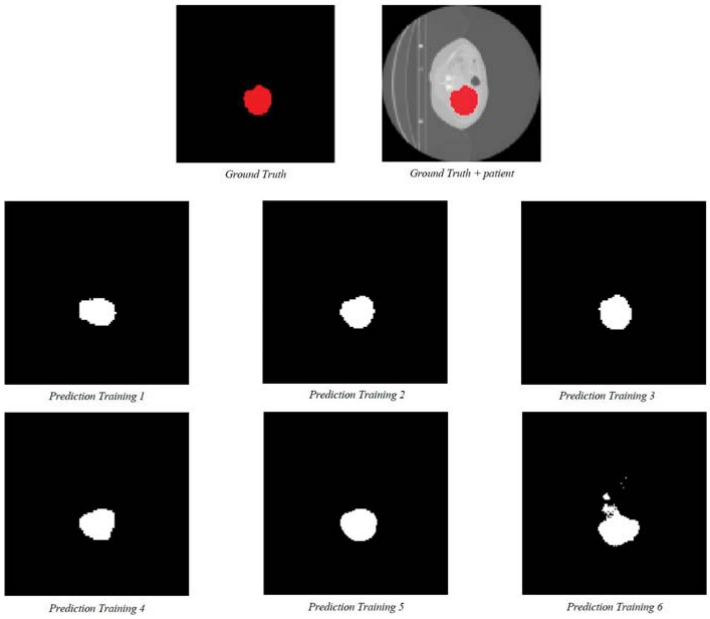

Figure 4: Visual representation of the UNETR results.

Citation: Tanya Fernández-Fernández, Lucía Cubero, Carmen Morote-García, Ana Álvarez González, Mercedes Muñóz-Fernández, Lydia Mediavilla-Santos, Rubén Pérez-Mañanes, Javier Pascau, José Antonio Calvo-Haro. Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications. Journal of Surgery and Research. 8 (2025): 274-284.

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

This image shows the ground truth segmentation of a pelvic bone sarcoma (top row) compared with UNETR model predictions from six training scenarios (bottom two rows). The ground truth and its overlay on the patient's CT scan are provided for reference.

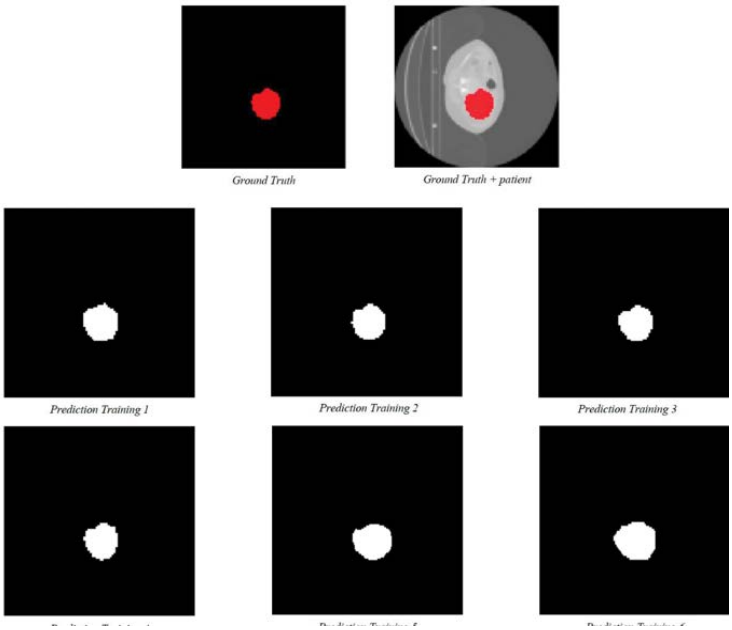

Figure 5: Visual representation of the SwinUNETR results.

The 3D volumetric representations of the tumors further illustrate the models' capacity to comprehend the full spatial extent of the tumor within the CT scans (Figure 6). These visualizations underscore the importance of evaluating models not only through quantitative metrics but also through qualitative assessment to ensure clinical relevance and applicability.

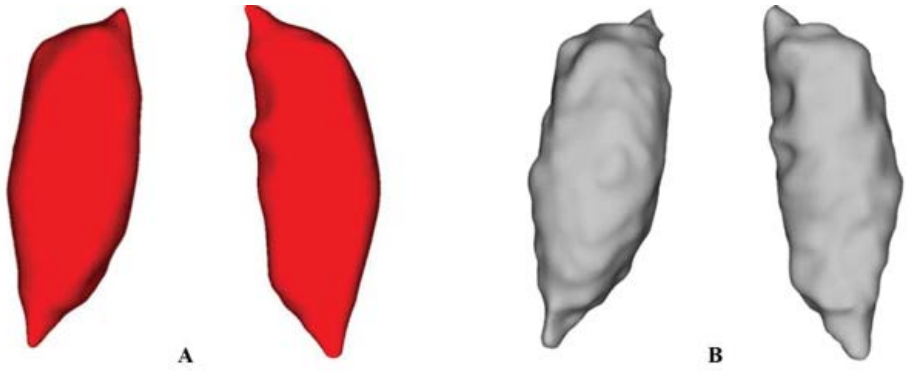

Figure 6: Visual 3D representation of the original (red) and predicted (grey) pelvic bone sarcomas.
A: Ground truth. **B:** Prediction from this dataset's best model.

Impact of training conditions

Throughout the experiments, several patterns were observed regarding the impact of different training conditions on model performance (Table 5). Increasing the batch size generally led to improved Dice scores, whereas the inclusion of the background in the Dice metric stabilized the performance. However, increasing the image size and introducing data augmentation during preprocessing sometimes results in decreased performance, highlighting the importance of careful consideration of these factors in model training. The use of cropped patient images had mixed results, and no clear conclusions could be drawn regarding its impact.

This table summarizes the effects of various training modifications on the performance of four different CNN-based frameworks used for tumor segmentation. Modifications such as increasing the batch size and the number of epochs generally improved Dice scores, although at the cost of longer training times. Conversely, increasing the image size and introducing data augmentation often results in decreased performance, highlighting the trade-offs inherent in optimizing model training across different architectures.

Citation: Tanya Fernández-Fernández, Lucía Cubero, Carmen Morote-García, Ana Álvarez González, Mercedes Muñóz-Fernández, Lydia Mediavilla-Santos, Rubén Pérez-Mañanes, Javier Pascau, José Antonio Calvo-Haro. Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications. Journal of Surgery and Research. 8 (2025): 274-284.

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

Table 5: Impact of training modifications on the performance of the four CNN-based frameworks.

Modification	Training time	Dice (test)
Increase batch size	↑	↑
Include background = True in the Dice Metric	~	↑
Increase number of epochs	↑	↑
Introduce Data Augmentation in preprocessing	↑	↓
Increase image size	↑	↓
Introduce cropped patient	~	-

Legend: ↑ increased, ↓ decreased, ~ no apparent effect, - conclusions could not be drawn

Discussion

Recent advancements in DL have significantly enhanced the state-of-the-art in medical image segmentation, with artificial neural networks (ANNs) at the forefront of these developments. Notably, segmentation strategies using DL algorithms have demonstrated a marked improvement in accuracy and robustness, particularly for challenging cases such as bone metastases and pelvic bone tumors [27]. These networks leverage the powerful feature extraction capabilities of convolutional neural networks (CNNs) and the ability to capture intricate spatial dependencies through transformer-based models [10]. As a result, these models have set new benchmarks in segmentation performance, often achieving DSCs well above the 0.7 commonly accepted threshold for adequate model performance; notably, many models now surpass this benchmark, achieving DSCs above0.85 in a few cases [28]. In the present study, we aimed to evaluate the effectiveness of four DL frameworks—U- Net, SegResNet, UNETR, and SwinUNETR—in segmenting pelvic bone tumors within a high-complexity hospital environment. This research holds particular significance due to its potential to enhance treatment planning and patient outcomes by automating the traditionally labor-intensive and variable process of tumor segmentation. Using the MONAI framework, we systematically explored these architectures to identify the most viable solutions for clinical implementation. Among the architectures evaluated, U-Net emerged as the most balanced, providing robust segmentation results with relatively low computational demands. Its best performance for pelvic bone sarcoma was a DSC of 81.788 ± 21.837, which was achieved with training times of just 15 minutes and 36 seconds. This efficiency and accuracy make U-Net particularly suitable for clinical settings with limited computational resources. SegResNet, however, faces challenges due to GPU memory constraints, limiting its ability to utilize data augmentation and larger image sizes. Despite these limitations, SegResNet demonstrated satisfactory performance, with a DSC of 81.241 ± 21.094 in some cases, although it required careful parameter tuning depending on the clinical conditions. UNETR, while theoretically advanced in capturing complex spatial dependencies, underperformed, with the best DSC of 81.664 ± 21.216. This suggests that UNETR may need further optimization or greater computational power to fully exploit its potential in clinical practice.SwinUNETR has emerged as the most promising architecture, outperforming UNETR in both accuracy and consistency. It achieved the highest DSC (82.080 ± 0.225), although this was achieved with longer training times, such as 38 minutes and 20 seconds. While SwinUNETR's superior performance is compelling, the increased computational cost could limit its adoption in settings where rapid model training is necessary. Despite these promising findings, several limitations of this study must be acknowledged. The limited GPU memory (12 GB of RAM) significantly constrained the size of the images that could be processed, potentially leading to the loss of critical information. This also limits the use of extensive data augmentation, potentially impacting model generalizability. Additionally, the restricted dataset size may have impacted the models' robustness, particularly given the variability in tumor shapes and locations. Generalizability remains a critical issue. While the models were evaluated in a specific clinical context, broader validation across different settings and tumor types is necessary [27]. The success of models trained on large public datasets [29] underscores the need for global efforts in this area. A recent study by Wu et al. (2024) introduced an advanced FCNN-4 s + CRF algorithm, which achieved superior DSCs- 91.000 (89.82, 92.57) -and real-time performance, highlighting the rapid advancements in this field [30]. Lastly, the present study did not explore the integration of more recent architectural advancements, such as transformer-based models beyond SwinUNETR, which may offer additional benefits in segmentation performance [25]. However, our study focuses on accessible segmentation tools that can be integrated into public hospital workflows, where large patient volumes and constrained computational resources are common challenges. In summary, this study underscores the importance of evaluating and optimizing accessible DL tools for clinical use, particularly in resource-constrained settings. U-Net and SwinUNETR stand out as the most feasible options for enhancing pelvic bone tumor segmentation efficiency. Future research should aim to enhance these models for broader clinical applications, ensuring scalability and generalizability across various tumor types. Integrating automated segmentation into routine clinical practice has the potential to significantly improve patient care and streamline radiotherapy and surgical planning processes.

Conclusions

Four DL frameworks —U-Net, SegResNet, UNETR, and SwinUNETR— have been evaluated for their ability to segment pelvic bone tumors within a high-complexity hospital setting. U-Net and SwinUNETR have emerged as the most clinically viable methods, as they balance

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

accuracy, efficiency, and resource use. Limitations, including GPU memory constraints and dataset size, affect model performance and generalizability. Despite these challenges, by identifying the most efficient frameworks for pelvic bone tumor segmentation, this study contributes to ongoing efforts to integrate automated segmentation into routine clinical workflows. The results provide a foundation for future research aimed at further optimizing these models to develop a generalized, scalable solution capable of accurately segmenting tumors in the pelvic region and beyond. This work underscores the potential of DL to transform medical imaging and improve patient care in oncology.

Conflicts of interest

The authors declare that they have no competing interests.

Acknowledgments

Analysis and interpretation of the data were supported by Projects TED2021-132200B- I00, TED2021-129392B-I00 and PID2023-149604OB-I00 (Ministerio de Ciencia e Innovación/AEI/10.13039/501100011033 and European Union "NextGenerationEU"/PRTR). This study has been funded by Instituto de Salud Carlos III (ISCIII) through the Biomodels and Biobanks Platform and co-funded by the European Union (PT23/00116 to RPM). We also acknowledge support from the PTI FAB3D, Consejo Superior de Investigaciones Científicas (CSIC), Spain.

References

1. Strauss SJ, Frezza AM, Abecassis N, et al. Bone sarcomas: ESMO-EURACAN-GENTURIS-ERN PaedCan Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 32 (2021): 1520-1536.
2. Puchner SE, Funovics PT, Böhler C, et al. Oncological and surgical outcome after treatment of pelvic sarcomas. PLoS One 12 (2017).
3. Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. Journal of Medical Physics/Association of Medical Physicists of India 33 (2008): 136.
4. Joskowicz L, Cohen D, Caplan N, et al. Interobserver variability of manual contour delineation of structures in CT. Eur Radiol 29 (2019): 1391-1399.
5. Covert EC, Fitzpatrick K, Mikell J, et al. Intra- and interoperator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry. EJNMMI Phys 9 (2022).
6. Chen Z, King W, Pearcey R, et al. The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. Radiother Oncol 87 (2008): 3-16.
7. Strachna O, Asan O. Reengineering Clinical Decision Support Systems for Artificial Intelligence. 2020 IEEE International Conference on Healthcare Informatics, ICHI (2020).
8. Rogowska J. Overview and Fundamentals of Medical Image Segmentation. Handbook of Medical Imaging. (2000): 69-85.
9. Shen D, Wu G, Suk H Il. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 19 (2017): 221-248.
10. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351 (2015): 234-241.
11. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 42 (2017): 60-88.
12. Mazurowski MA, Buda M, Saha A, et al. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. Journal of Magnetic Resonance Imaging 49 (2019): 939–954.
13. Harrison K, Pullen H, Welsh C, et al. Machine Learning for Auto-Segmentation in Radiotherapy Planning. Clin Oncol (R Coll Radiol) 34 (2022): 74-88.
14. Kalantar R, Lin G, Winfield JM, et al. Automatic segmentation of pelvic cancers using deep learning: State-of-the-art approaches and challenges. Diagnostics 11 (2021).
15. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. J Big Data 6 (2019).
16. Li X, Morgan PS, Ashburner J, et al. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods 264 (2016): 47-56.
17. NIfTI:-Neuroimaging Informatics Technology Initiative. https://nifti.nimh.nih.gov/. Accessed 17 Aug 2024.
18. Cardoso MJ, Li W, Brown R, et al. MONAI: An open-source framework for deep learning in healthcare 655 (2022).
19. MONAI - Home. https://monai.io/. Accessed 17 Aug 2024.
20. 3D Slicer image computing platform | 3D Slicer. https://www.slicer.org/. Accessed 17 Aug 2024.
21. Joseph VR. Optimal ratio for data splitting. Stat Anal Data Min 15 (2022): 531-538.
22. Zhao R, Qian B, Zhang X, et al. Rethinking Dice Loss for Medical Image Segmentation. Industrial Conference on Data Mining (2020): 851-860.

Fernández-Fernández T, et al., J Surg Res 2025
DOI:10.26502/jsr.10020449

23. Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation 669 (2016).

24. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D Medical Image Segmentation. IEEE Workshop/Winter Conference on Applications of Computer Vision (2021): 1748-1758.

25. Hatamizadeh A, Nath V, Tang Y, et al. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. ArXiv. 12962 LNCS (2022): 272-284.

26. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging 15 (2015): 1-28.

27. Paranavithana IR, Stirling D, Ros M, et al. Systematic Review of Tumor Segmentation Strategies for Bone Metastases. Cancers 15 (2023).

28. Rich JM, Bhardwaj LN, Shah A, et al. Deep learning image segmentation approaches for malignant bone lesions: a systematic review and meta-analysis. Frontiers in Radiology 3 (2023).

29. Wu J, Yang S, Gou F, et al. Intelligent Segmentation Medical Assistance System for MRI Images of Osteosarcoma in Developing Countries. Comput Math Methods Med (2022).

30. Wu S, Ke Z, Cai L, et al. Pelvic bone tumor segmentation fusion algorithm based on fully convolutional neural network and conditional random field. J Bone Oncol 45 (2024).

Citation: Tanya Fernández-Fernández, Lucía Cubero, Carmen Morote-García, Ana Álvarez González, Mercedes Muñóz-Fernández, Lydia Mediavilla-Santos, Rubén Pérez-Mañanes, Javier Pascau, José Antonio Calvo-Haro. Evaluating deep learning models for Pelvic Bone Tumor Segmentation: Implications for Radiotherapy and Surgical Applications. Journal of Surgery and Research. 8 (2025): 274-284.