Original Article

Mol Imaging Radionucl Ther **DOI:***10.4274/mirt.galenos.2026.56578*

Comparison of Classical Image Processing Methods and the Artificial Intelligence-Based MedSAM2 Model for Liver Tumor Segmentation in Radioembolization Dosimetry: A LiTS Dataset Analysis

Radyoembolizasyon Dozimetrisinde Karaciğer Tümörü Segmentasyonu: Klasik Görüntü İşleme Yöntemleri ile Yapay Zeka Tabanlı MedSAM2 Modelinin LiTS Veri Setinde Karşılaştırılması

Özlem Maranci Yıldırım

Altınbaş University, Vocational School of Health Services, Medical Imaging Techniques Program, İstanbul, Türkiye

Abstract

Objectives: Accurate estimation of tumor and critical organ doses in radioembolization therapy is essential for treatment planning. The choice of segmentation technique used in voxel-based dosimetry can affect the calculated volumes and dose estimates. In this study, the liver tumor segmentation performances of classical image processing methods and a pretrained artificial intelligence (AI) Medical Segment Anything Model 2 (MedSAM2) were compared, along with their effects on a simplified dose-change indicator.

Methods: From the 131 training cases in the Liver Tumor Segmentation Benchmark dataset, 33 eligible cases were identified using previously published radioembolization eligibility criteria; of these, 30 patients were selected by tumor-volume-stratified sampling. In each patient, four classical methods [dilation, erosion, region of interest (ROI)-based Otsu thresholding, ROI-based K-means clustering] and one AI-based method (MedSAM2) were applied to the reference tumor mask, and the resulting volumes were compared with the reference volume to calculate a simplified medical internal radiation dosimetry-derived dose-change percentage.

Results: All five methods were successfully completed in all 30 patients. Mean±standard deviation (median) dose-change percentages were -25.5%±8.9% (-25.3%) for dilation, 50.9%±30.2% (43.1%) for erosion, 177.3%±1324.8% (-78.3%) for ROI-thresholding, -65.0%±33.0% (-79.0%) for ROI-K-means, and -30.2%±40.7% (-34.6%) for MedSAM2. ROI-thresholding produced an extreme deviation (7189.2%) in one patient due to a degenerate threshold; MedSAM2 showed a similar extreme deviation (-99.2%) in one patient by segmenting an area approximately 123-fold the reference volume. Excluding these outliers, MedSAM2's variance remained lower than that of the classical thresholding and clustering methods.

Conclusion: The choice of liver tumor segmentation method can lead to differences ranging from 10% to over 1000% in simplified dose estimates. Classical thresholding and clustering methods tend to produce infrequent but severe outliers, whereas a pretrained AI model (MedSAM2) yields more stable results, although it is not entirely error-free. These findings point to the importance of standardizing and reporting the segmentation method used in pre-treatment dosimetric assessment for radioembolization.

Keywords: Radioembolization, dosimetry, liver neoplasms, computer-assisted image processing, artificial intelligence

Address for Correspondence: Özlem Maranci Yıldırım, Altınbaş University, Vocational School of Health Services, Medical Imaging Techniques Program, İstanbul, Türkiye

E-mail: ozlem.yildirim1@altinbas.edu.tr **ORCID ID:** orcid.org/0000-0002-0749-8388

Received: 24.07.2026 **Accepted:** 11.10.2026 **Epub:** 25.11.2026

Cite this article as: Maranci Yıldırım Ö. Comparison of classical image processing methods and the artificial intelligence-based MedSAM2 model for liver tumor segmentation in radioembolization dosimetry: a LiTS dataset analysis. Mol Imaging Radionucl Ther. [Epub Ahead of Print]

Copyright© 2027 The Author(s). Published by Galenos Publishing House on behalf of the Turkish Society of Nuclear Medicine.
This is an open access article under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

Öz

Amaç: Radyoembolizasyon tedavisinde tümör ve kritik organ dozlarının doğru tahmini, tedavi planlamasının güvenilirliği için kritik önem taşır. Voksel-tabanlı dozimetri yöntemlerinde kullanılan segmentasyon tekniğinin seçimi, hesaplanan hacimleri ve dolayısıyla doz tahminlerini doğrudan etkileyebilir. Bu çalışmada, klasik görüntü işleme yöntemleri ile hazır eğitilmiş bir yapay zeka (YZ) modelinin (MedSAM2) karaciğer tümörü segmentasyon performansları ve bunların basitleştirilmiş doz değişimi üzerindeki etkileri karşılaştırılmıştır.

Yöntem: Liver Tumor Segmentation Benchmark veri setinin 131 eğitim olgusundan, önceden yayımlanmış klinik radyoembolizasyon uygunluk kriterlerine göre 33 uygun olgu belirlenmiş, tümör hacmine göre tabakalı örnekleme ile 30 hasta seçilmiştir. Her hastada referans tümör maskesine dört klasik yöntem [morfolojik genişletme, morfolojik daraltma, lokal ilgi alanı (ROI) tabanlı Otsu eşiklemesi, lokal ROI tabanlı K-means kümeleme] ve bir YZ tabanlı yöntem (MedSAM2) uygulanmış, elde edilen hacimler referans hacimle karşılaştırılarak basitleştirilmiş tıbbi iç radyasyon dozimetrisi türevi bir doz değişim yüzdesi hesaplanmıştır.

Bulgular: Otuz hastanın tamamında beş yöntem başarıyla tamamlanmıştır. Doz değişim yüzdesinin ortalama ± standart sapma (medyan) değerleri sırasıyla dilation için -%25,5±%8,9 (-%25,3), erosion için %50,9±%30,2 (%43,1), ROI-eşikleme için %177,3±%1324,8 (-%78,3), ROI-K-means için -%65,0±%33,0 (-%79,0) ve MedSAM2 için -%30,2±%40,7 (-%34,6) olarak bulunmuştur. ROI-eşikleme bir hastada dejenere bir eşik değeri nedeniyle aşırı bir sapma (%7189) üretmiş; MedSAM2 ise bir hastada referans hacmin ~123 katı bir alanı segmentleyerek benzer bir aşırı sapma (-%99,2) göstermiştir. Bu tekil aşırı değerler hariç tutulduğunda MedSAM2'nin varyansı, klasik eşikleme/kümeleme yöntemlerine kıyasla belirgin şekilde daha düşük kalmıştır.

Sonuç: Karaciğer tümörü segmentasyonunda yöntem seçimi, basitleştirilmiş doz tahminlerinde %10'dan %1000'in üzerine varan farklara yol açabilmektedir. Klasik eşikleme ve kümeleme yöntemleri nadir ama şiddetli aşırı değerler üretme eğilimindedir; hazır eğitilmiş bir YZ modeli (MedSAM2) genel olarak daha kararlı sonuçlar vermekle birlikte tamamen hatasız değildir. Bulgular, radyoembolizasyon öncesi dozimetrik değerlendirmede segmentasyon yönteminin standartlaştırılması ve şeffaf biçimde raporlanmasının önemine işaret etmektedir.

Anahtar Kelimeler: Radyoembolizasyon, dozimetri, karaciğer neoplazmları, bilgisayar destekli görüntü işleme, yapay zeka

Introduction

Radioembolization (transarterial radioembolization, selective internal radiation therapy) is an intra-arterial brachytherapy technique based on the administration of microspheres labeled with a beta-emitting radioisotope (most commonly yttrium-90, ^{90}Y). It is used to treat advanced primary and metastatic liver tumors, particularly in cases that are resistant to chemotherapy or when chemoembolization cannot be performed because of portal vein occlusion. Safe and effective delivery of treatment depends on accurate pre-treatment estimation of the absorbed dose to the tumor and surrounding critical organs (liver parenchyma, lungs, stomach, spleen, kidneys) (1).

Clinically, dose estimation is generally performed using simplified, mono-compartmental models derived from the medical internal radiation dosimetry (MIRD) formalism; these models assume homogeneous activity distribution within the organ of interest and disregard patient-specific anatomical variations (2,3). As an alternative, increasingly common in the literature, patient-specific voxel-based Monte Carlo (MC) dosimetry approaches (4,5,6) rely on accurate segmentation of tumor and organ volumes from computed tomography (CT) images as the first and most critical step of the calculation chain; the author's previous study also showed that MCNP-based MC simulations could estimate critical organ doses in greater detail than MIRD (2).

Errors in the segmentation step or inconsistencies between methods are directly reflected in the calculated volumes and, consequently, in the absorbed dose, which depends on mass (m) [via the simplified MIRD relationship D=50×A/m). However, few studies have systematically examined the effect of volume differences produced by different segmentation methods on dose estimation in the same patient. Deep learning-based approaches (e.g., a deep neural network method proposed for whole-body voxel dosimetry (7)) have also shown promising results in segmentation/dose estimation, but studies directly comparing these methods with classical image processing techniques are scarce.

In the field of liver and liver tumor segmentation, the publicly shared dataset of 131 training cases created as part of the Liver Tumor Segmentation Benchmark (LiTS) challenge, held alongside the 2017 IEEE ISBI and 2017-2018 MICCAI conferences (8), provides an important resource for comparing segmentation methods under standard and reproducible conditions. The reference labels (liver and tumor) provided with this dataset allow different methods to be objectively evaluated against the same reference.

Classical segmentation methods (thresholding, clustering, morphological operations, region growing, watershed) are simple and fast, but are sensitive to parameter selection and to the histogram characteristics of the image. In recent years, segmentation models trained with the "foundation model" approach and steerable by prompts – the Segment Anything Model (SAM) and its medical-imaging-adapted derivative, Medical Segment Anything Model 2 (MedSAM2) (9) – have emerged as an alternative approach; these models can propagate segmentation from a single 2D slice, given a box or point prompt, to an entire three-dimensional

structure via a "video tracking" mechanism. No official and validated pretrained weight set compatible with the currently used nnU-Net v2 framework for the LiTS/Liver task could be found for fully supervised deep learning models such as nnU-Net (10), which is still considered the reference method in liver tumor segmentation (only weights in the older nnU-Net v1 format, not directly compatible with the current framework, are available); therefore, MedSAM2, an accessible and validated pretrained model, was preferred in this study.

The aim of this study is to comparatively evaluate, within the framework of the project titled "Effect of Different Image Processing Methods on Tumor and Critical Organ Doses for Radioembolization Patients" (No: PB2025-SHMYO-1), which is supported by the Altınbaş University Scientific Research Projects (BAP) unit, the tumor-volume estimates produced by four classical segmentation methods and by a pretrained AI model (MedSAM2) in 30 patients selected from the LiTS dataset, and their effects on a simplified MIRD-derived dose change. The findings constitute a preliminary assessment for the project's next work package, in which each segmentation method will be compared with full MCNP-based MC dosimetry in a subset of 20 patients. They also provide guidance on which methods are reliable for this comparison and which should be interpreted with caution because of high variability.

Materials and Methods

Dataset and Patient Selection

The 131 publicly available training cases from the LiTS dataset were used (8). Each case contains a CT volume (volume-X.nii) and its corresponding reference segmentation label (segmentation-X.nii; 0=background, 1=liver, 2=tumor), both in NIfTI-1 format.

Four criteria were used to select patients. The first was adapted from the clinical radioembolization eligibility criterion defined in the author's previous study (2): (i) the ratio of tumor volume to total liver volume (liver+tumor) being below 50%. In addition, three technical criteria specific to this study were applied: (ii) a minimum tumor volume of ≥1 cm^3; (iii) the number of discrete (non-connected) tumor foci in a patient being between 1 and 3, i.e., excluding extremely multifocal/miliary disease with numerous scattered small tumor foci; (iv) an inter-slice z-axis spacing on CT of ≤2.5 mm (ensuring adequate image resolution). Of the 33 patients meeting these four criteria, 30 were selected using stratified sampling based on tumor-volume tertiles (small/medium/large); this approach prevented the study from being concentrated in a single tumor-size group. The selection flow is summarized in Figure 1.

Reference Segmentation and Volume Calculation

For each patient, the reference tumor volume was calculated as the number of tumor voxels in the reference label (label=2) multiplied by the voxel volume (pixel spacing × pixel spacing × slice spacing).

Segmentation Methods Compared

Five different methods were applied on the reference tumor mask or within the local region of interest (ROI) it defines:

(1) Morphological dilation: One iteration of morphological dilation was applied to the reference tumor mask using a 6-connected neighborhood structure (considering only the six face-adjacent neighbors of a voxel, i.e., the neighboring voxels in the anterior-posterior, left-right, and superior-inferior directions) (11).

(2) Morphological erosion: One iteration of morphological erosion was applied using the same 6-connected neighborhood structure (11).

(3) ROI-based Otsu thresholding: Within a local ROI defined by adding an 8-voxel margin to the bounding box of the reference tumor, an automatic threshold value was determined using Otsu's method (12) from the HU values of the liver voxels (label ≥1) inside this ROI; of the two voxel

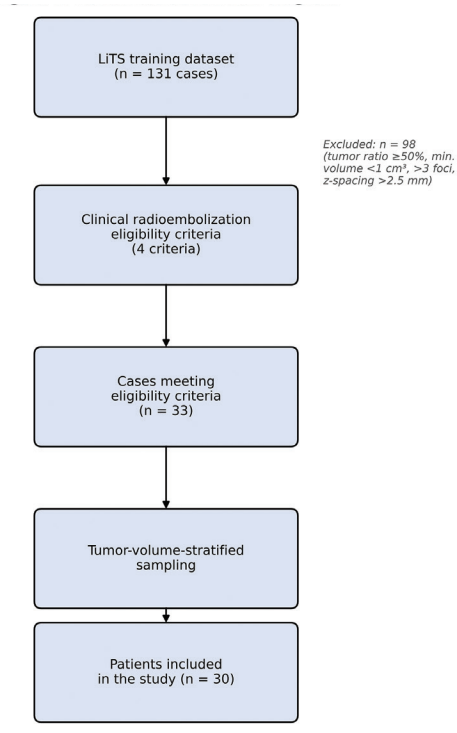


Figure 1. Selection flow of the 30-patient segmentation study subset from the LiTS dataset
LiTS: Liver Tumor Segmentation Benchmark

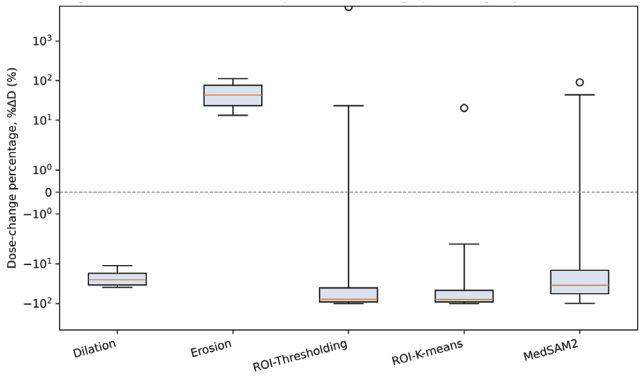


Figure 2. Distribution of the simplified dose-change percentage (%ΔD) by method (symmetric-log y-axis; circles indicate statistical outliers)

sets separated by this threshold, the one containing fewer voxels was labelled as tumor (the rationale for this minority rule is given under Method 4).

(4) ROI-based K-means clustering: One-dimensional K-means clustering with k=2 (13) was applied to the Hounsfield unit (HU) values of the liver voxels (label ≥1) within the same local ROI, and the minority cluster was accepted as tumor. This cardinality-based assignment rule was adopted for two reasons. First, because the ROI is the bounding box of the reference tumor enlarged by an 8-voxel margin in all three directions, and clustering is restricted to liver voxels, parenchyma necessarily outnumbers tumor tissue: across the 30 patients, the reference tumor accounted for a median of 8.8% (range 0.3-29.2%) of the liver voxels inside the ROI and was the minority class in all 30 patients. Second, the rule does not presuppose that the lesion is hypodense relative to parenchyma, an assumption that does not hold uniformly in this dataset; in 4 of the 30 patients, the minority cluster was the higher-HU cluster. Replacing the minority rule with a fixed hypodensity rule (always assigning the lower-HU cluster to tumor) was tested post hoc and changed the mean Dice of the method only marginally, from 0.294 to 0.285, with individual changes in both directions (largest gain, from 0.040 to 0.489; largest loss, from 0.318 to 0.166). The minority rule was therefore retained as the assignment independent of the direction of the intensity difference and applicable to both hypodense and hyperdense lesions; all results reported in this study were produced using this rule. The same rule was used to label the two classes produced by Otsu thresholding (method 3), so methods 3 and 4 differ only in how the boundary between the two classes is determined, not in how the resulting classes are labelled.

(5) MedSAM2 (9): The MedSAM2_CTLesion.pt weights, shared in the bowang-lab/MedSAM2 repository and specifically trained for CT lesions, were used. This model is a derivative of the SAM architecture, adapted to video and 3D medical images. In each patient, a 2D bounding-box prompt (±5-pixel margin), automatically extracted from the middle axial slice of the reference tumor's z-axis, was given to the model; starting from this single slice, the model propagated the segmentation both forward and backward via the video-tracking (propagation) mechanism to the reference tumor's local z-range (±8-voxel margin). The largest connected component of the resulting three-dimensional mask was taken as the final MedSAM2 segmentation. Because the box prompt was automatically derived from the reference mask, the evaluation was semi-automatic; this point is addressed as a limitation in the Discussion section.

Software and Computing Environment

All analyses were performed using Python 3.14.6 (64-bit) on Windows 11 (build 10.0.26200). The four classical methods, the NIfTI-1 reader, and the three-dimensional connected-component analysis were implemented in custom code using NumPy 2.4.4 as the only numerical dependency and without an external image-processing library. Figures were produced with Matplotlib 3.11.1. The one-dimensional K-means implementation used k=2, five random initialisations (n_init=5), at most 50 iterations per initialisation and a fixed random seed (seed=0); the classical pipeline is therefore deterministic, and re-running it reproduced every reported classical value identically.

MedSAM2 inference was performed with PyTorch 2.11.0+cu128 and torchvision 0.26.0+cu128 (CUDA 12.8) on an NVIDIA GeForce RTX 5060 GPU. The inference code was obtained from the bowang-lab/MedSAM2 repository (distributed package version 1.0, commit 332f30d420f1d1b08e2a79b3ae6a602458808383), which was installed in editable mode; hydra-core 1.3.4 and iopath 0.1.10 were used for model configuration and checkpoint loading, and nibabel 5.4.2, pandas 3.0.3, and scikit-image 0.26.0 were used for image input/output and connected-component labelling. The model configuration file was configs/sam2.1_hiera_t512.yaml. The model weights were stored in MedSAM2_CTLesion.pt, downloaded on 21 July 2026 from the Hugging Face repository wanglab/MedSAM2 (revision e4a6f35edd7e091619cbc0750f462f1574e23955; file size 156,041,079 bytes; SHA-256 78f7e125418df d6fec22f4afe90bcd85cb1d4423d0a9df36f7a87ed63 aa1a5f5).

Simplified Dose Change

For each method, the calculation was based on the simplified dose relationship of the MIRD formalism:

$$D = 50 \times A/m \quad (1)$$

where D is the absorbed dose (Gy), A is the activity (GBq) and m is the mass (kg). Under the assumption that activity (A) remains constant while mass (m) is directly proportional to volume, a dose-change percentage (%ΔD) relative to the reference volume (Vref) was calculated:

$$\%\Delta D = (V_{ref}/V_{method} - 1) \times 100 \quad (2)$$

Equation 2 is not an actual MIRD dose calculation; rather, it is a simplified indicator used solely to illustrate, in comparative terms, the order of magnitude of the potential dose impact resulting from volume differences between segmentation methods.

Statistical Analysis

For each method, the mean, median, standard deviation (SD), minimum, and maximum of the dose-change percentages across the 30 patients were calculated. Outliers were defined as cases in which a method produced a volume markedly >5-fold or <0.2-fold the reference volume; this symmetric (±5-fold on a logarithmic scale) threshold is based on a fold-change convention commonly used for outlier screening in ratio data, and corresponds to approximately -80% to +400% in terms of the simplified dose-change indicator; as there is no consensus clinical cut-off in the literature on this matter, the threshold was defined arbitrarily but symmetrically and transparently. In these cases, both the mean and the median were reported; when outliers markedly skewed the mean, the median was emphasized as the more representative summary statistic. In addition to the dose-change percentage, the spatial overlap accuracy of each method with the reference mask was evaluated using the Dice similarity coefficient (14) and the Jaccard index (15); the use of these two metrics in 3D medical image segmentation evaluation is discussed in (16). This calculation was performed for all methods across all patients (n=30). The overall difference among the dose-change percentages of the five methods was assessed for statistical significance using the Friedman test (17), and pairwise comparisons were assessed using the Bonferroni-corrected Wilcoxon signed-rank test (18); the significance level was set at α=0.05.

Results

All five methods were successfully completed for the 30 selected patients (LiTS dataset case numbers); the patient list is provided in the study's data files. Summary statistics for the dose-change percentage, by method, are presented in Table 1, and their distribution is shown in Figure 2.

Morphological dilation, as expected, systematically increased tumor volume, producing a consistent negative dose change (mean -25.50%) and showing the lowest variability (SD=8.88%); morphological erosion, in contrast, consistently reduced volume, producing a positive dose change (mean 50.85%). The predictable and monotonic nature of these two methods is also supported by their low SDs.

ROI-based Otsu thresholding selected a degenerate threshold value (-211 HU) in one patient (segmentation-52), segmenting only 1.4% of the reference volume (0.25 cm^3) (Figure 3B). This produced a meaningless dose change of 7189.24%. This single outlier completely distorts the method's mean (177.25%), whereas its median (-78.27%) is more representative. When this single case is excluded (n=29), the mean decreases to -64.55% and the SD decreases from 1324.79% to 34.77%, which illustrates how dramatically a single outlier can dominate summary statistics.

MedSAM2 produced reasonable results, within 0.2- to 5-fold of the reference volume, in 29 of the 30 patients (a typical case is shown in Figure 3A). However, in one patient

Table 1. Summary statistics of the simplified dose-change percentage (%ΔD) by segmentation method

Method	Mean (%)	Median (%)	SD (%)	Min (%)	Max (%)	n
Dilation (morphological dilation)	-25.50	-25.32	8.88	-39.20	-11.14	30
Erosion (morphological erosion)	50.85	43.12	30.15	13.45	112.16	30
ROI-based thresholding (Otsu)	177.25	-78.27	1324.79	-99.41	7189.24	30
ROI-based thresholding (excluding segmentation-52)	-64.55	-79.11	34.77	-99.41	23.44	29
ROI-based K-means	-65.00	-78.98	33.04	-99.44	20.35	30
MedSAM2	-30.24	-34.60	40.65	-99.18	90.46	30
MedSAM2 (excluding segmentation-42)	-27.86	-32.83	39.19	-91.57	90.46	29

%ΔD = (Vref/Vmethod -1) × 100, where Vref is the reference tumor volume and Vmethod is the volume obtained with the given method. For the two methods containing one marked single outlier (ROI-based thresholding and MedSAM2), sensitivity-analysis results excluding the relevant outlier case are also presented in separate rows. No patient was excluded from the study; all statistics are calculated over n=30, and the outlier-excluded rows are provided only as an additional sensitivity check. SD: Standard deviation, ROI: Region of interest, MedSAM2: Medical Segment Anything Model 2

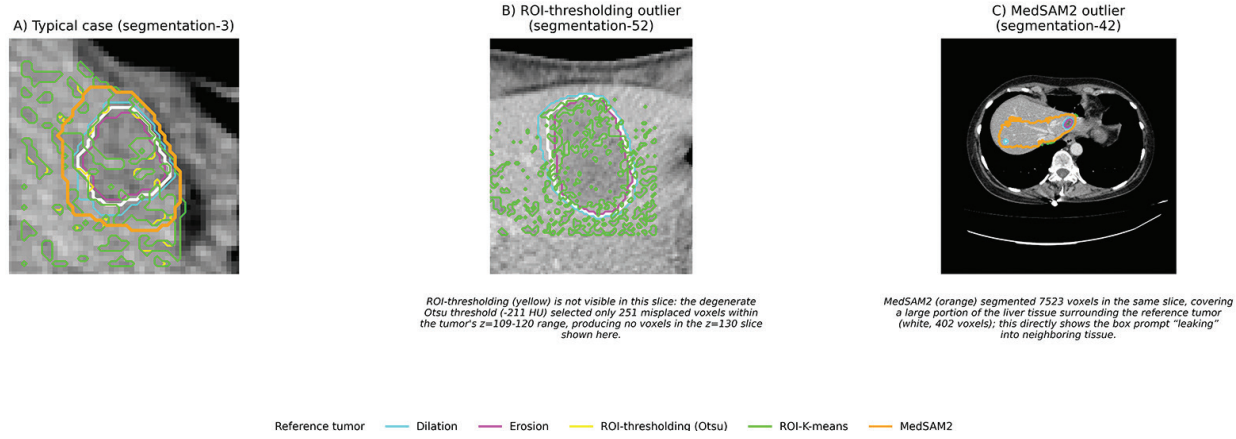


Figure 3. (A) Comparison of the reference tumor contour with the four classical methods and MedSAM2 in a typical case (segmentation-3). (B) A case in which ROI-based thresholding produced a degenerate result (segmentation-52): the yellow contour is not visible in this slice because the erroneous segmentation is located in a different z-range of the tumor (see figure note). (C) A case in which MedSAM2 produced over-segmentation (segmentation-42). In the wide-angle view, MedSAM2 (orange) can be seen extending far beyond the reference tumor (white), covering a large portion of the adjacent liver tissue

ROI: Region of interest, MedSAM2: Medical Segment Anything Model 2

(segmentation-42), it segmented a region approximately 123-fold the reference volume (200.97 cm^3 vs. 1.64 cm^3), constituting a severe outlier (-99.18%) (Figure 3C). This is thought to result from the box prompt "leaking" into a neighboring anatomical structure. When this single case is excluded, MedSAM2's variability (SD=39.19%) is much lower than that of ROI-based thresholding (SD=1324.79%) and similar to that of K-means (SD=33.04%); however, MedSAM2's median dose change (-32.83%) is negative, indicating that the model tends to systematically predict volumes somewhat larger than the reference–likely because of the box prompt's (+5-pixel) margin and the tendency for forward and backward video propagation to expand tumor boundaries.

Because the dose-change indicator is based on a volume ratio, two methods producing the same volume at different locations could yield similar %ΔD values, creating a misleading impression of similarity. Therefore, the spatial overlap between each method and the reference mask was evaluated using the Dice similarity coefficient and the Jaccard index (Table 2).

The Dice and Jaccard results confirm and reinforce the variance finding obtained from the dose-change percentage: dilation and erosion, which are small morphological modifications of the reference mask, showed the highest overlap, as expected (mean Dice values of 0.851 and 0.808, respectively). In contrast, the mean Dice values of ROI-based Otsu thresholding and K-means, which derive the segmentation solely from image intensity within the local ROI, were markedly lower (0.280 and 0.294, respectively) and showed a very wide distribution (Dice≈0 in some patients); this directly demonstrates that these methods are inconsistent not only in dose estimation but also in spatial segmentation accuracy. MedSAM2, with Dice values calculated across all 30 patients (mean ± SD: 0.636±0.220; median: 0.704), showed markedly higher overlap performance than the classical thresholding/clustering methods, but lower than the morphological reference methods (dilation, erosion); the Jaccard index reflects the same pattern (mean ± SD: 0.500±0.215; median: 0.543). That the median is markedly higher than the mean (0.704 vs. 0.636) indicates that the distribution is skewed downward by a small number of low-Dice cases; indeed, the lowest Dice values were observed in the same case that also showed an extreme deviation in the dose-change analysis (segmentation-42, Dice=0.016) and in a few other patients (e.g., segmentation-55, Dice=0.153). These findings indicate that MedSAM2 provides a more reliable spatial overlap than the classical thresholding and clustering methods in most patients, but is less consistent than the morphological reference methods and continues to produce serious segmentation errors in a small number of patients.

The overall difference among the dose-change percentages of the five methods was assessed with the Friedman test and was found to be significant ($\chi 2(4)=76.55$; $p<0.001$). In Bonferroni-corrected pairwise Wilcoxon signed-rank tests, differences were statistically significant (corrected $p<0.05$) for the following pairs: dilation-erosion, dilation-K-means, erosion-thresholding, erosion-K-means, erosion-MedSAM2, and K-means-MedSAM2; no significant difference was found between dilation-MedSAM2 and thresholding-

Table 2. Spatial overlap accuracy (Dice, Jaccard) of segmentation methods versus the reference mask

Method	Dice (mean ± SD)	Dice median	Jaccard (mean ± SD)	Jaccard median	n
Dilation (morphological dilation)	0.851±0.058	0.855	0.745±0.089	0.747	30
Erosion (morphological erosion)	0.808±0.093	0.823	0.688±0.129	0.699	30
ROI-based thresholding (Otsu)	0.280±0.243	0.239	0.189±0.196	0.136	30
ROI-based K-means	0.294±0.240	0.244	0.198±0.193	0.139	30
MedSAM2	0.636±0.220	0.704	0.500±0.215	0.543	30

SD: Standard deviation, ROI: Region of interest, MedSAM2: Medical Segment Anything Model 2

MedSAM2. These results, which are consistent with the descriptive statistics (Table 1), confirm that the difference between methods is not due to chance. However, this test only evaluates differences in the central tendency (median) of the dose change and does not directly assess differences in variance or stability between methods.

Discussion

The comparison of four classical image processing methods with a pretrained AI model (MedSAM2) for liver tumor segmentation in 30 patients selected from the LiTS dataset showed that the choice of segmentation method can lead to differences in simplified dose estimates ranging from -99% to +7189%. This finding indicates that the segmentation step is a far more critical source of uncertainty in pre-radioembolization voxel-based dosimetry than is often assumed.

The predictable monotonic bias shown by the morphological methods (dilation/erosion) is expected, and these methods may be useful only for establishing "best-case/worst-case" bounds; their direct use in clinical dosimetry is not recommended. The extreme variance shown by histogram-based methods such as ROI-based thresholding and K-means (particularly SD=1324.79% for thresholding) results from these methods' high sensitivity to image contrast, noise, and the HU similarity between the lesion and surrounding tissue; this suggests that classical, parameter-dependent segmentation methods may be risky to use as the sole reliable reference in clinical dosimetry.

MedSAM2's generally more stable behavior (SD=39.19% when the single outlier is excluded) can be regarded as an advantage of foundation-model-based approaches over classical methods. However, the single but severe outlier observed (segmentation-42) shows that these models are not error-free and that careful quality control is needed, particularly in automatic/semi-automatic box-prompt scenarios. In this study, the box prompt was automatically derived from the reference mask, which somewhat separates MedSAM2's performance from that in a real clinical-use scenario (a prompt manually drawn by a user). Therefore, the results reported here may be closer to the model's "best-case" performance; repeated evaluation with genuine user-derived prompts should be conducted in future studies.

These findings directly inform the next stage planned within the Altınbaş University BAP project (PB2025-SHMYO-1): MC (MCNP6.3)-based voxel dosimetry will be performed on a subset of 20 patients, prioritizing the methods found to be more stable in this study.

Study Limitations

This study has important limitations. First, LiTS is a publicly available CT dataset and does not include the actual ^{90}Y activity distribution, single photon emission computed tomography/positron emission tomography images, or clinical follow-up outcomes of real radioembolization patients. Therefore, the simplified dose-change indicator used here is not a substitute for an actual MIRD- or MC-based dose calculation and only provides a comparative illustration of possible order of magnitude of segmentation-related volume differences. Second, the scope of this study was deliberately limited to approaches that do not require training (the four classical methods) and that can be used directly with pretrained weights (MedSAM2); the aim is to compare methods that clinical or medical physics teams can apply directly, without requiring additional model-training infrastructure. nnU-Net (10), still considered the reference/best-performing method in the literature for liver tumor segmentation, was not included in this comparison because it requires dataset-specific training from scratch, which would require a separate training/validation process beyond the scope of the current study (the absence of an official pretrained LiTS/Liver weight set directly compatible with the currently used nnU-Net v2 framework further supported this choice); comparison with fully supervised, from-scratch-trained state-of-the-art methods (including nnU-Net) is planned as an explicit goal of a future study. Third, the limited sample size (n=30) and reliance on a single public dataset restrict the generalizability of the findings.

Further studies with larger patient cohorts and, where possible, different data sources will reduce this limitation.

Conclusion

The choice of liver tumor segmentation method significantly affects volume estimates (and, consequently, simplified dose estimates) used in radioembolization dosimetry. Classical thresholding and clustering methods can produce rare but severe outliers, whereas a pretrained AI model (MedSAM2) generally offers more stable results, although it is not entirely error-free. These findings point to the importance of clearly reporting the segmentation method used in pre-treatment dosimetric assessment for radioembolization and, where possible, cross-validating it with more than one method.

Ethics

Ethics Committee Approval: This study used the publicly available, fully anonymized Liver Tumor Segmentation Benchmark dataset; no new patient data were collected.

Informed Consent: This is a retrospective study based on a publicly available, fully anonymized dataset; no new patients were recruited and no new data were collected, so informed consent was not applicable.

Artificial Intelligence (AI) Use: Claude (Anthropic, Claude Sonnet 5) was used to assist with drafting and English-language translation of the manuscript text; writing and verifying the Python code used for image processing, statistical analysis, and figure generation; and formatting the manuscript to meet journal requirements. All AI-assisted text, code, and analyses were reviewed, validated, and approved by the author prior to submission.

Data and Code Availability: The Liver Tumor Segmentation Benchmark dataset used in this study is publicly available at https://www.kaggle.com/datasets/harshwardhanbhangale/lits-dataset. The custom Python code written for the segmentation methods, volume/dose-change calculations, and statistical analyses is retained in the corresponding author's project files and can be provided upon reasonable request.

Footnotes

Authorship Contributions

Concept: Ö.M.Y., Design: Ö.M.Y., Data Collection or Processing: Ö.M.Y., Analysis or Interpretation: Ö.M.Y., Literature Search: Ö.M.Y., Writing: Ö.M.Y.

Financial Disclosure: This research project was supported by Altınbaş University Scientific Research Fund (project number: PB2025-SHMYO-1).

References

1. Kennedy A, Nag S, Salem R, Murthy R, McEwan AJ, Nutting C, Benson A 3rd, Espat J, Bilbao JI, Sharma RA, Thomas JP, Coldwell D. Recommendations for radioembolization of hepatic malignancies using yttrium-90 microsphere brachytherapy: a consensus panel report from the radioembolization brachytherapy oncology consortium. Int J Radiat Oncol Biol Phys. 2007;68:13-23.
2. Yildirim OM, Bingolbali A, Bozkurt A, Cafaro C, Demirkol MO. Monte Carlo calculations of critical organ doses in radioembolization therapy of primary liver tumors via ^{90}Y microspheres. Int J Radiat Res. 2023;21:353-360.
3. Loevinger R, Budinger TF, Watson EE. MIRD Primer for Absorbed Dose Calculations. Rev ed. New York: Society of Nuclear Medicine; 1991.
4. Chiavassa S, Aubineau-Lanièce I, Bitar A, Lisbona A, Barbet J, Franck D, Jourdain JR, Bardiès M. Validation of a personalized dosimetric evaluation tool (Oedipe) for targeted radiotherapy based on the Monte Carlo MCNPX code. Phys Med Biol. 2006;51:601-616.
5. Petitguillaume A, Bernardini M, Hadid L, de Labriolle-Vaylet C, Franck D, Desbrée A. Three-dimensional personalized Monte Carlo dosimetry in ^{90}Y resin microspheres therapy of hepatic metastases: nontumoral liver and lungs radiation protection considerations and treatment planning optimization. J Nucl Med. 2014;55:405-413.
6. Gulec SA, Sztejnberg ML, Siegel JA, Jevremovic T, Stabin M. Hepatic structural dosimetry in ^{90}Y microsphere treatment: a Monte Carlo modeling approach based on lobular microanatomy. J Nucl Med. 2010;51:301-310.
7. Akhavanallaf A, Shiri I, Arabi H, Zaidi H. Whole-body voxel-based internal dosimetry using deep learning. Eur J Nucl Med Mol Imaging. 2021;48:670-682.
8. Bilic P, Christ P, Li HB, Vorontsov E, Ben-Cohen A, Kaissis G, Szeskin A, Jacobs C, Mamani GEH, Chartrand G, Lohöfer F, Holch JW, Sommer W, Hofmann F, Hostettler A, Lev-Cohain N, Drozdzal M, Amitai MM, Vivanti R, Sosna J, Ezhov I, Sekuboyina A, Navarro F, Kofler F, Paetzold JC, Shit S, Hu X, Lipková J, Rempfler M, Piraud M, Kirschke J, Wiestler B, Zhang Z, Hülsemeyer C, Beetz M, Ettlinger F, Antonelli M, Bae W, Bellver M, Bi L, Chen H, Chlebus G, Dam EB, Dou Q, Fu CW, Georgescu B, Giró-I-Nieto X, Gruen F, Han X, Heng PA, Hesser J, Moltz JH, Igel C, Isensee F, Jäger P, Jia F, Kaluva KC, Khened M, Kim I, Kim JH, Kim S, Kohl S, Konopczynski T, Kori A, Krishnamurthi G, Li F, Li H, Li J, Li X, Lowengrub J, Ma J, Maier-Hein K, Maninis KK, Meine H, Merhof D, Pai A, Perslev M, Petersen J, Pont-Tuset J, Qi J, Qi X, Rippel O, Roth K, Sarasua I, Schenk A, Shen Z, Torres J, Wachinger C, Wang C, Weninger L, Wu J, Xu D, Yang X, Yu SC, Yuan Y, Yue M, Zhang L, Cardoso J, Bakas S, Braren R, Heinemann V, Pal C, Tang A, Kadoury S, Soler L, van Ginneken B, Greenspan H, Joskowicz L, Menze B. The liver tumor segmentation benchmark (LiTS). Med Image Anal. 2023;84:102680.
9. Ma J, Yang Z, Kim S, Chen B, Baharoon M, Fallahpour A, Asakereh R, Lyu H, Wang B. MedSAM2: segment anything in 3D medical images and videos. arXiv:2504.03600 [Preprint]. 2025.
10. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203-211.
11. Serra J. Image Analysis and mathematical morphology. London: Academic Press; 1982.

12. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9:62-66.
13. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1967. p. 281-297.
14. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26:297-302.
15. Jaccard P. The distribution of the flora in the alpine zone. New Phytol. 1912;11:37-50.
16. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29.
17. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32:675-701.
18. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1:80-83.