Radiotherapy and Oncology 197 (2024) 110338

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Development and benchmarking of a Deep Learning-based MRI-guided gross tumor segmentation algorithm for Radiomics analyses in extremity soft tissue sarcomas

Jan C. Peeken [a,b,c,d,1], Lucas Etzel [a,b,1,*], Tim Tomov [e], Stefan Münch [a], Lars Schüttrumpf [a], Julius H. Shaktour [a], Johannes Kiechle [e], Carolin Knebel [f], Stephanie K. Schaub [g], Nina A. Mayr [h], Henry C. Woodruff [d,i], Philippe Lambin [d,i], Alexandra S. Gersing [j], Denise Bernhardt [a], Matthew J. Nyflot [g,k], Bjoern Menze [e], Stephanie E. Combs [a,b,c], Fernando Navarro [e]

[a] Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich (TUM), Munich, Germany
[b] German Consortium for Translational Cancer Research (DKTK), Partner Site Munich, Munich, Germany
[c] Institute of Radiation Medicine (IRM), Helmholtz Zentrum München (HMGU), German Research Center for Environmental Health GmbH, Neuherberg, Germany
[d] Department of Precision Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University, Maastricht, Netherlands
[e] Department of Informatics, Technical University of Munich (TUM), Garching, Germany
[f] Department of Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical University of Munich (TUM), Munich, Germany
[g] Department of Radiation Oncology, University of Washington, Seattle, USA
[h] College of Human Medicine, Michigan State University, East Lansing, MI, USA
[i] Department of Radiology and Nuclear Imaging, GROW - School for Oncology and Developmental Biology, Maastricht University Medical Centre, the Netherlands
[j] Institute of Neuroradiology, LMU Klinikum, LMU Munich, Munich, Germany
[k] Department of Radiology, University of Washington, Seattle, USA

ARTICLE INFO

Keywords:
Soft tissue sarcoma
Deep Learning
Radiomics
MRI
Radiotherapy
Radiology
Tumor Volume

ABSTRACT

Background: Volume of interest (VOI) segmentation is a crucial step for Radiomics analyses and radiotherapy (RT) treatment planning. Because it can be time-consuming and subject to inter-observer variability, we developed and tested a Deep Learning-based automatic segmentation (DLBAS) algorithm to reproducibly predict the primary gross tumor as VOI for Radiomics analyses in extremity soft tissue sarcomas (STS).
Methods: A DLBAS algorithm was trained on a cohort of 157 patients and externally tested on an independent cohort of 87 patients using contrast-enhanced MRI. Manual tumor delineations by a radiation oncologist served as ground truths (GTs). A benchmark study with 20 cases from the test cohort compared the DLBAS predictions against manual VOI segmentations of two residents (ERs) and clinical delineations of two radiation oncologists (ROs). The ROs rated DLBAS predictions regarding their direct applicability.
Results: The DLBAS achieved a median dice similarity coefficient (DSC) of 0.88 against the GTs in the entire test cohort (interquartile range (IQR): 0.11) and a median DSC of 0.89 (IQR 0.07) and 0.82 (IQR 0.10) in comparison to ERs and ROs, respectively. Radiomics feature stability was high with a median intraclass correlation coefficient of 0.97, 0.95 and 0.94 for GTs, ERs, and ROs, respectively. DLBAS predictions were deemed clinically suitable by the two ROs in 35% and 20% of cases, respectively.
Conclusion: The results demonstrate that the DLBAS algorithm provides reproducible VOI predictions for radiomics feature extraction. Variability remains regarding direct clinical applicability of predictions for RT treatment planning.

* Corresponding author at: Department of Radiation Oncology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich (TUM), Ismaninger Straße 22, 81675 Munich, Germany.
E-mail address: lucas.etzel@tum.de (L. Etzel).
[1] These authors contributed equally, shared first authorships.

https://doi.org/10.1016/j.radonc.2024.110338
Received 30 July 2023; Received in revised form 5 May 2024; Accepted 10 May 2024
Available online 22 May 2024

Introduction

Extremity soft tissue sarcomas (STS) constitute a rare and challenging malignancy. Postoperative and preoperative radiation therapy (RT), which requires meticulous characterization and delineation of the tumor, are standard of care in high grade (G2 and G3) STS [1–4]. Imaging data derived from computed tomography (CT) or magnetic resonance imaging (MRI) are a crucial information source for the treatment planning process [4,5].

The planning target volumes for RT are based on accurate volume of interest (VOI) segmentations of the STS gross tumor volume (GTV). Time-consuming manual delineation still represents the pivotal step in the radiation oncology workflow. The process of defining the optimal GTV for STS remains challenging because of their complex and unfamiliar imaging anatomy within musculoskeletal and soft tissue structures. Further, the presumed true GTV delineation can vary between different readers, as it may depend on different interpretations of tumor infiltration or edema in surrounding tissues [6,7]. Thus, GTV segmentation may be susceptible to inter-observer variability even in highly trained readers. A more objective approach would greatly improve robustness and reproducibility of tumor delineation in STS.

In recent years, Deep Learning (DL) and Radiomics-based machine learning have emerged as research areas within the scope of artificial intelligence (AI) as novel methods for extracting additional information from medical imaging such as CT or MRI [8]. Several studies have shown these methods to be useful for prognostic assessment and tumor characterization in patients with STS [9–13].

Besides being essential for the treatment planning process, VOI segmentation is also important for Radiomics analyses by providing the basis for feature extraction or bounding box generation for DL applications. The quality of these VOI definitions is a relevant factor for radiomic feature stability [14]. Recent advances using DL-based algorithms such as the U-Net may also reduce the time needed for VOI segmentation [15]. The use of such an algorithm additionally bears the potential to improve segmentation reproducibility and inter-observer variability.

The purpose of this work was to develop a Deep Learning-based automatic segmentation (DLBAS) algorithm for VOI prediction of extremity STS based on fat-saturated contrast-enhanced T1-weighted (T1-CE) MRI sequences. The final model derived from the training cohort was tested using an external patient cohort. The stability of the radiomic features extracted from the DL-based VOI predictions was assessed. In an additional benchmark study, the value of VOI prediction was assessed in a sub-cohort of the test cohort by evaluating the segmentation performance in comparison to manual and DL-assisted segmentations. The time taken for segmentation as well as the direct clinical applicability of VOI predictions as GTV delineations for RT treatment planning were also assessed.

Methods

Patient cohorts

Based on T1-CE MRI sequences, we used a training cohort of 157 patients from the University of Washington/Fred Hutchinson Cancer Center (UW/FHCC) to develop a U-Net-based algorithm for STS VOI segmentation. An independent cohort of 87 patients from the Technical University of Munich (TUM) was used as a test cohort.

Volume of interest definition

With the purpose of radiomic feature extraction and neural network development, the VOI was defined as the clearly delineable contrast-enhanced tumor bulk. Unclear areas of infiltration such as diffuse contrast enhancement in surrounding tissues that may be included into a GTV delineation for RT treatment planning were deliberately excluded. Manual VOI delineations of all MRI studies were performed by a board-certified radiation oncologist (author JCP) to serve as ground truth segmentations (GTs).

Data preprocessing

Data standardization was performed on the imaging data prior to model development and inference. All volumes were resampled to 1 mm isotropic resolution and normalized using z-score normalization. During training of the algorithm, we cropped a region of interest around the GT segmentation using a randomized distance to the tumor boundary between 5–10 voxels in every direction. Using a randomized distance increases the variability in the bounding box selection, thus aiming at improving robustness of the segmentation algorithm. The coordinates for the cropping origin were manually input into the network, thus yielding a semi-automatic procedure.

Neural network development

The segmentation model is a modified version of the 3D U-Net architecture, consisting of an encoder part acting as feature extractor, and a decoder part bringing back the features to input image resolution. Horizontal connections are used to pass information between encoder and decoder at the different feature levels, as depicted in Fig. 1.

We added squeeze and excitation blocks in the convolutions according to Roy et al. [16], as well as residual blocks [17] to aid gradient flow and avoid vanishing gradients. Furthermore, we included multi-head self-attention at the bottleneck of the network to enforce the network to look for spatial relationships that would be limited when only using convolutional windows [18]. The modified 3D U-Net was determined empirically during training.

Optimization of Deep Learning models

All models were developed in Pytorch with a 12 GB Titan XP [19]. The models were trained with a batch size of 1 and a learning rate of $1x10^{-4}$ with an ADAM optimizer for 100 epochs, using early stopping during training on the validation loss to avoid overfitting. A combination of Dice-loss and binary cross-entropy was used as a loss function. The Dice-loss was used to mitigate the class imbalance between the foreground and background.

Network evaluation strategy

The performance, reproducibility and generalizability of the algorithm was evaluated using stratified 5-fold cross-validation. During inference, the final segmentation output of the model is an ensemble of the 5-fold, averaging the soft-max activation function from the 5-folds and computing the maximum between the foreground and background class.

Radiomic feature extraction

In total 104 radiomic features were extracted as previously described [11]. Further details regarding image discretization [40] and radiomics feature extraction [41] are described in the supplemental materials. See **Supplemental Table S1** for all features.

Segmentation benchmark

In a subgroup composed of patients from the external test cohort, we performed a benchmark study: For each imaging study, two first-year residents (“early residents,” ERs, authors LE and JHS) manually performed VOI segmentations and modified the respective DLBAS predictions following the segmentation approach used for definition of the GTs. Furthermore, two board-certified radiation oncologists (ROs,

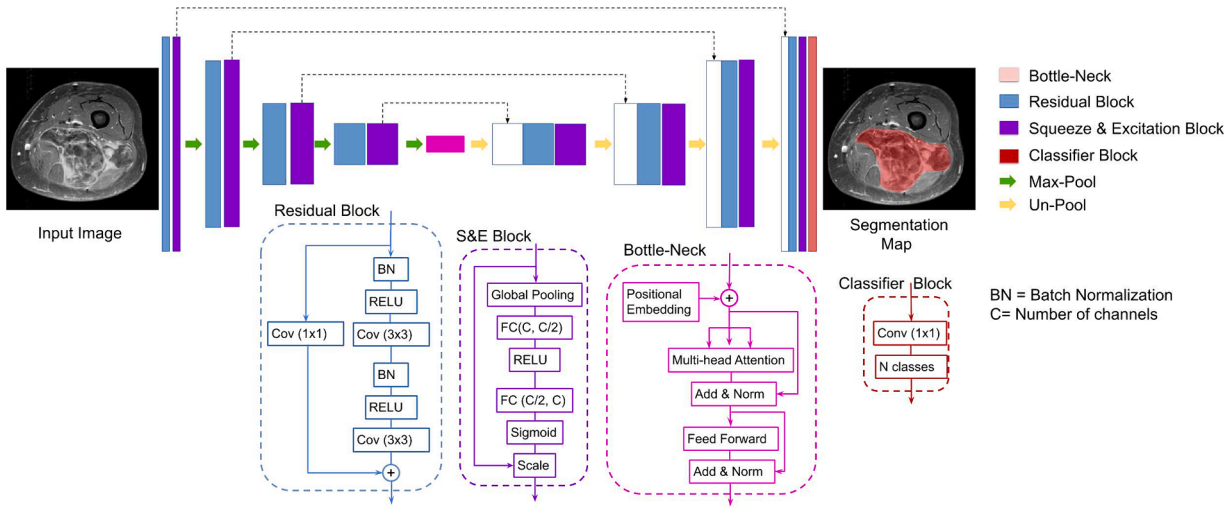

Fig. 1. Deep Learning segmentation network architecture.

authors SM and LS) manually defined segmentations. Independent of the radiation oncologist who defined the GT segmentations, these ROs delineated the primary gross tumor in a clinical approach including diffuse contrast enhancement extending beyond the tumor bulk as clinically indicated for RT treatment planning. The ROs also modified the DLBAS predictions using the same clinical approach. The purpose of examining these differing delineation methodologies between GT/ERs and ROs was to gauge the differences between clinically suited gross tumor volumes for the treatment planning process, in comparison to our VOIs that were developed specifically for radiomics feature extraction.

Sample size calculation

The total number of test subjects was chosen based on a sample size calculation. We calculated a necessary sample size of 19 subjects to be able to detect a moderate effect size (Cohen's d 0.6) for a paired one-tailed *t*-test with a significance level of 0.05 and a power of 0.8 [20]. We rounded the sample size to 20 patients.

Segmentation procedures

Prior to segmentation, the MRI studies and DLBAS prediction label maps were transformed from their respective source file formats (".nrrd" and ".nii.gz") into Digital Information and Communications in Medicine (DICOM) file formats using the open-source Plastimatch software package [21]. The resulting DICOM files were then imported into the segmentation software. Segmentation for the benchmark study was performed in Eclipse 13.0 (Varian Medical Systems, Palo Alto, CA, USA) using two different segmentation methods:

Manual segmentation: This method involved manual VOI segmentation from scratch using all available drawing tool functions.

Deep Learning-assisted segmentation: This method involved assessment of the VOI segmentation generated by the DL network and subsequent manual correction of the label map using a drawing tool function as deemed necessary by the operator. The operators were told to adapt all automatic segmentations, including those that would have been directly clinically applicable.

Segmentation randomization

To reduce the case-specific learning effect of repetitive segmentations, both segmentation methods were conducted in a randomized manner in two sessions with a four-week interval. The imaging studies were split in a cross-over design, so that each operator used both methods in each segmentation session. The segmentation sequence was randomized individually for each operator.

Radiomic feature stability assessment

The radiomic feature stability was assessed using ICC (3,1; two-way mixed effect) by comparing the features extracted from the DLBAS predictions with the features extracted from the segmentations defined manually by the ERs and ROs. Substudy analyses were conducted to examine feature stability dependent on varying shape feature values of the DLBAS predictions. For this, the benchmark test set was stratified based on the median values of the Flatness, Elongation, Sphericity and Surface/Volume-Ratio shape features, thus yielding two respective datasets of high and low values for further ICC analyses. We also tested a previously trained radiomics model for overall survival prediction using the DLBAS predictions [10]. Due to different inclusion and exclusion criteria the test patient cohort from the original publication was reduced from 71 to 64 patients.

Segmentation performance assessment

To analyze segmentation performance, VOI segmentations were compared with the original GTs by calculating the dice similarity coefficients (DSC) as well as the Hausdorff distance (HD). After quantitative assessments of inter-observer variability between the performance metric samples within each group of ERs and ROs, the data samples were pooled for ERs and ROs, respectively.

Assessment of spatial deviances

The spatial deviance between the DLBAS prediction and the VOIs defined by the ERs or ROs were graphically analyzed using agreement maps.

Segmentation time assessment

The time taken for image segmentation was measured using a digital stopwatch from the beginning of each VOI segmentation process to its conclusion. The segmentation times were compared between both segmentation methods while pooling the data for both pairs of ERs and ROs.

Need for manual correction

To assess whether the DLBAS predictions may be directly applicable

as GTV delineations for the RT planning process, the frequency of a need for manual correction was noted during segmentation. For this purpose, a binary value (correction necessary vs. correction not necessary) was assigned to each DLBAS task by each of the respective ROs.

Statistical analysis

Data transformation and analysis were conducted using the R software environment for statistical computing and graphics (version 4.1.2) [22] within the RStudio [23] IDE. We used R Markdown [24], R Bookdown [25] and the tidyverse collection of R packages [26]. Sample normality was assessed using the Shapiro-Wilk test, histograms, and qq-plots [27,28]. The Wilcoxon signed-rank test was used for hypothesis testing [28]. The irr package [29] was used to calculate ICC. A p-value <0.05 was considered statistically significant.

Results

The patient characteristics of the subject cohorts are shown in Table 1. Histologies are presented in **Table S2**.

When comparing the inter-rater variability of the ERs, the comparison of their respective manual segmentations yielded a median DSC of 0.92 (IQR: 0.86–0.94) and a median HD of 9.0 (IQR: 15.3–18.9).

For the ROs, the comparison of their respective manual segmentations yielded a median DSC of 0.88 (IQR: 0.84–0.90) and a median HD of 15.4 (IQR: 11.0–27.8).

Table 1
Patient cohort characteristics.

Characteristic	Detail	UW Training	TUM Testing	p-value
Total patients		n = 157	n = 87	p < 0.001
Median OS		49.2 months	43.2 months	p = 0.535
Age	Mean	53.7 years	56.3 years	p = 1.000
Age	Standard deviation	16.1 years	18 years	
Location	Lower extremity	108 (68.8%)	60 (69.0%)	p = 0.753
Location	Upper extremity	28 (17.8%)	13 (14.9%)	
Location	Pelvis	21 (13.4%)	14 (16.1%)	
Sex	Female	71 (45.2%)	39 (44.8%)	p = 1.000
Sex	Male	86 (54.8%)	48 (55.2%)	
T stage	1	28 (17.8%)	13 (14.9%)	p = 0.689
T stage	2	129 (82.2%)	74 (85.1%)	
T stage suffix	a	7 (4.5%)	7 (8.0%)	p = 0.386
T stage suffix	b	150 (95.5%)	80 (92.0%)	
N stage	0	157 (100.0%)	84 (96.6%)	p = 0.083
N stage	1	0 (0.0%)	3 (3.4%)	
M stage	0	152 (96.8%)	81 (93.1%)	p = 0.309
M stage	1	5 (3.2%)	6 (6.9%)	
Grading	1	27 (17.2%)	6 (6.9%)	p = 0.082
Grading	2	50 (31.8%)	32 (36.8%)	
Grading	2	80 (51.0%)	48 (55.2%)	
AJCC stage	IA	3 (1.9%)	1 (1.1%)	p = 0.001
AJCC stage	IB	13 (8.3%)	5 (5.7%)	
AJCC stage	IIA	28 (17.8%)	10 (11.5%)	
AJCC stage	IIB	33 (21.0%)	4 (4.6%)	
AJCC stage	III	75 (47.8%)	61 (70.1%)	
AJCC stage	IV	5 (3.2%)	0 (0.0%)	

Abbreviations: TUM: Technical University of Munich, UW: University of Washington. The listed p-values were determined for each grouped characteristic by the Chi-squared test.

When examining the performance of the segmentation algorithm, the DLBAS achieved a median DSC of 0.88 (IQR: 0.81–0.92) against the GTs for the entire test cohort and 0.87 (IQR: 0.84–0.92) for the benchmark cohort. The median HD values were 12.0 mm (IQR: 8.7–17.0) and 12.3 mm (IQR: 8.3–19.1), respectively.

In comparison to the manual segmentations of ERs and ROs, the DLBAS yielded a median DSC of 0.89 (IQR: 0.85–0.92) and 0.82 (IQR: 0.77–0.87), respectively. There was a statistically significant difference between these samples ($p < 0.001$).

The comparison of the ERs and ROs against the DLBAS yielded a median HD of 9.9 mm (IQR: 6.6–17.5) and 20.3 mm (IQR: 12.4–32.6), respectively. There was a statistically significant difference between these samples ($p < 0.001$).

A boxplot with DSC and HD results as performance metrics for the segmentation comparisons within the benchmark cohort is shown in Fig. 2. A boxplot with DSC and HD results as performance metrics for the segmentation comparisons of ERs and ROs with the GTs is shown in **Fig. S1**.

We compared the radiomic features extracted from the GTs of all imaging studies in the test set with the DL-based VOI predictions. The features showed a high stability with a median ICC of 0.97 (interquartile range (IQR) 0.93–0.98). A total of 98/104 features (94%) showed an ICC greater than 0.8.

We then examined the subset of the 20 imaging studies included in the benchmark study: The radiomic feature stability between the manual segmentations of the ERs and DL-based VOI predictions was also high with a median ICC of 0.95 (IQR 0.93–0.97). The radiomic feature stability between the manual segmentations of the ROs and DL-based VOI predictions was similar with a median ICC of 0.94 (IQR 0.88–0.97). Here, a total of 89/104 features (86%) showed an ICC greater than 0.8. The results from the ICC substudy analyses examining the feature stability dependent on varying shape feature values of the DLBAS predictions are shown in Tables S4, S5 and S6. Here, smaller Elongation and Flatness values yielded higher ICC values across all radiomic features. In contrast, greater values for the Surface/Volume-Ratio and Sphericity yielded similarly high ICC values. Furthermore, the median ICC values were no lower than 0.92 for the GTs/DL-based VOI predictions, no lower than 0.89 for the ROs/DL-based VOI predictions, and no lower than 0.89 for the ERs DL-based VOI predictions.

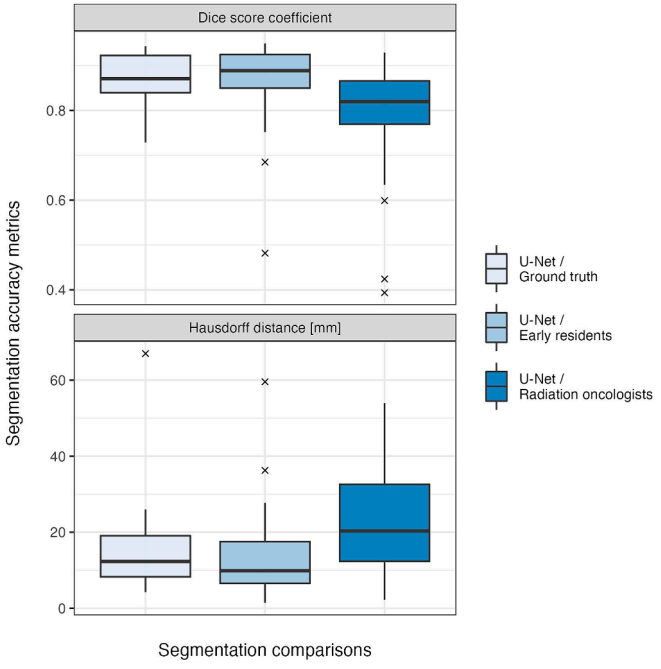

Fig. 2. Boxplot of segmentation performance metrics in the benchmark cohort.

The radiomics-based model based on the automated contours achieved a concordance index of 0.59 (0.44–0.73) while the manual contours achieved a similar concordance index of 0.60 (0.45–0.74).

The spatial deviance between the DLBAS predictions and the respective manually defined segmentations of each pair of ERs and ROs were graphically assessed. Agreement maps were generated on a slice-by-slice basis in the transverse plane for each MRI scan.

Key findings from the agreement maps of the ERs are addressed in the three rows of imaging studies depicted in Fig. 3. The agreement maps illustrate the challenge in clearly defining gross tumor boundaries in the presence of diffuse contrast enhancement in surrounding tissues.

The key findings from the agreement maps of the ROs are addressed in the three rows of imaging studies depicted in Fig. 4. In contrast to the DL-based VOI predictions, the ROs included areas of diffuse contrast enhancement in their clinical approach to target delineation.

When manually defining the VOI segmentations within the benchmark subgroup, the ERs required a median time of 6.1 min (IQR: 3.4–9.9), while the ROs required a median time of 3.7 min (IQR: 2.7–5.3). When modifying the DL-predicted VOI segmentations, the ERs required a median time of 6.4 min (IQR: 2.2–11.7), while the ROs required a median time of 4.3 min (IQR: 0.5–6.3). The comparison of the time taken for segmentation between ERs and ROs within the benchmark cohort is depicted in Fig. 5. There was a statistically significant difference in the time spent for manual segmentation between the ERs and ROs ($p = 0.002$).

The timing comparison between the first session and the second session for ERs and ROs within the benchmark cohort yielded no significant difference ($p = 0.747$ and $p = 0.347$) and is shown in **Supplemental Table S3**.

The two ROs rated the DL-based VOI segmentations as directly applicable for the clinical RT planning process as a clinical GTV in 7/20 (35%) cases and 4/20 (20%) cases, respectively.

Discussion

In this study, we developed a DL-based segmentation algorithm using a limited medical imaging dataset to predict VOI segmentations of extremity STS for Radiomics analyses.

There is a need for accurate and reproducible VOI segmentations to effectively extract stable radiomics features and apply these features for prognostic modeling [30–33]. To avoid the challenges and variability associated with manual segmentations [34], deep learning-based automatic segmentation has proven to be a powerful tool for radiomics feature extraction [35].

The algorithm presented in this article provided reproducible VOI predictions with good accordance compared to GT delineations from a radiation oncologist. When examining the ICC analysis results, the stability of the radiomic features extracted from these segmentations was high. Substudy analyses yielded good stability independent of selected shape features of the VOI predictions.

Both the radiation oncologist contouring the initial GTs as well as the ERs conducting VOI segmentations within the benchmark study focused on delineating the clearly definable gross tumor bulk while deliberately excluding unsure areas of potential tumor infiltration. These segmentations were intended to reduce inter-observer dependence and bias in radiomic feature extraction from the partial inclusion of healthy tissue with diffuse contrast-enhancement or from partial volume effects. Thus, the procedure was applied to generate VOIs as a basis for Radiomics-

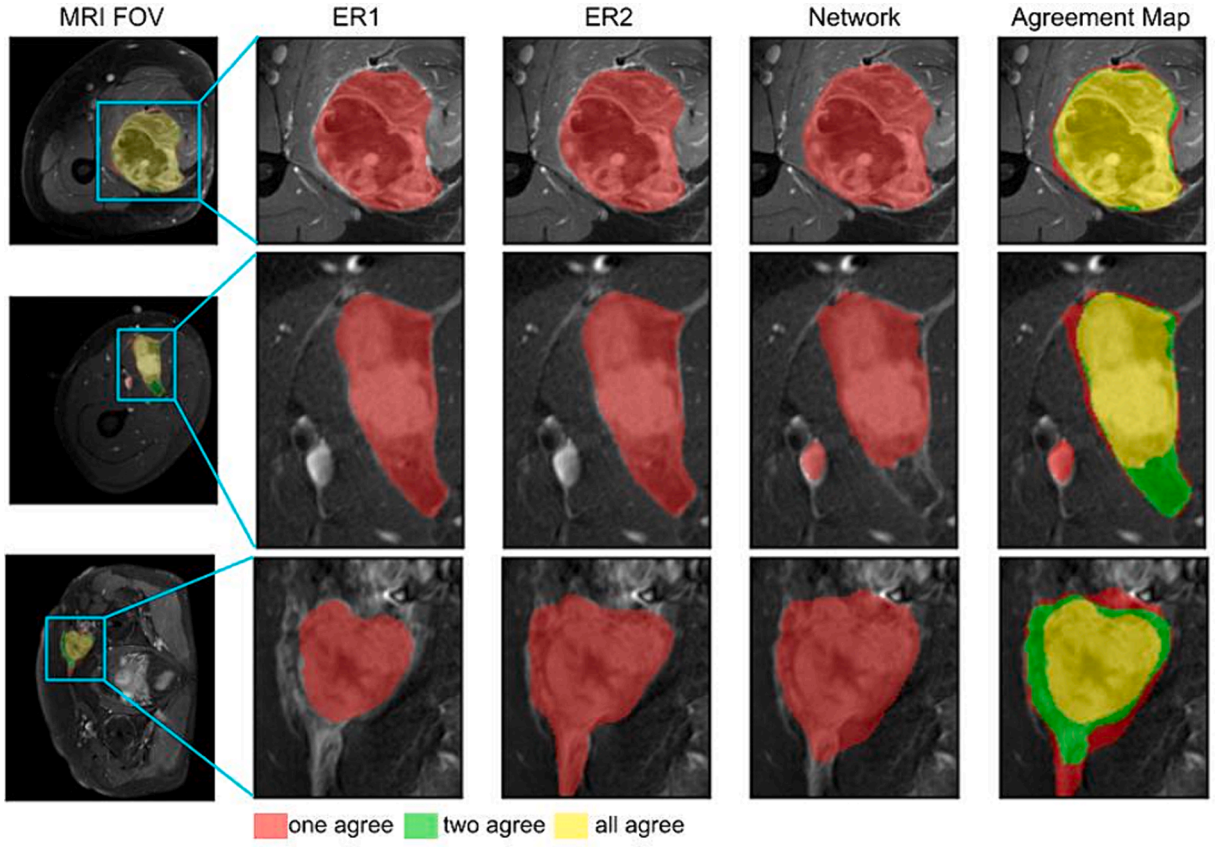

Fig. 3. Agreement maps for ERs and the DL network. **Row 1**: The GTV segmentations are similar between both ERs and the DL Network and show good visual congruence. **Row 2**: The DL Network prediction underestimated the muscular anatomical compartment that ERs included in their segmentation. Moreover, it falsely identified contrast enhancement such as blood vessels as GTV. **Row 3**: Both ERs and U-Net show the difficulties and subsequent discrepancies between label maps in cases with significant diffuse contrast enhancement extending beyond the tumor bulk. Whereas ER1 defined a conservative GTV, ER2 included most contrast. enhancement. The DL network shows some of both of these segmentation characteristics.

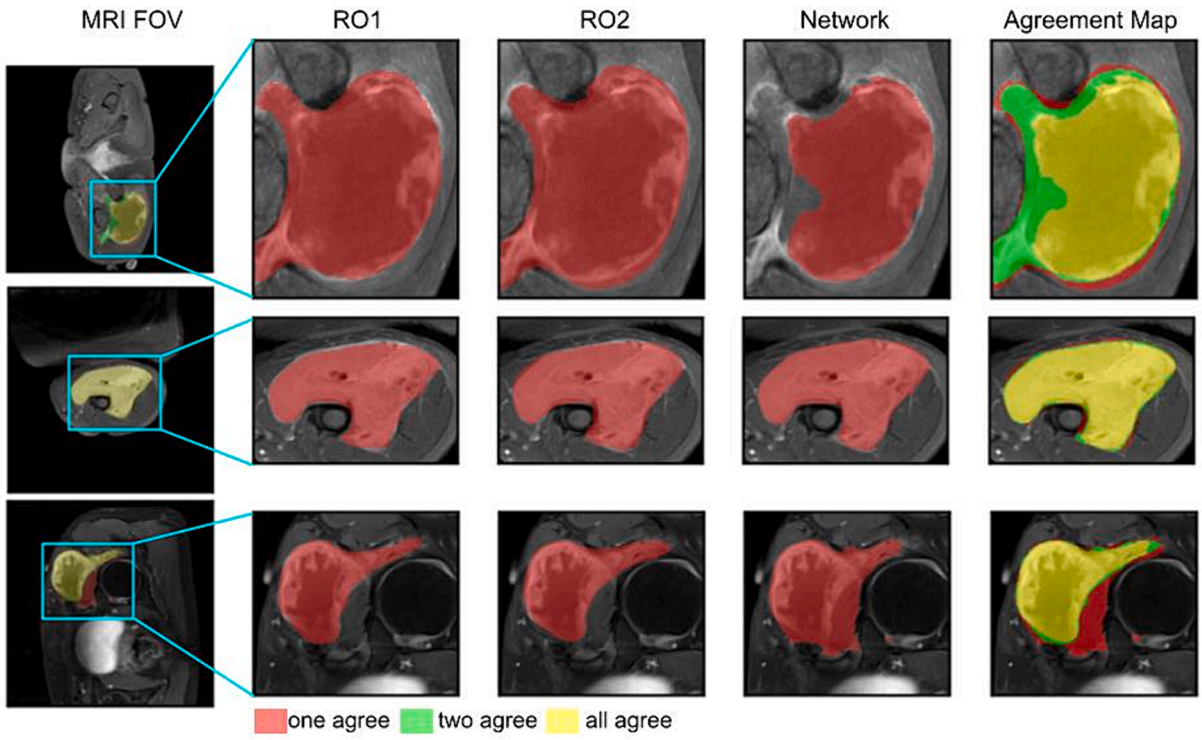

Fig. 4. Agreement maps for ROs and the DL network. **Row 1**: The DL Network focused on the tumor bulk and did not segment areas of diffuse contrast enhancement that were included by the two ROs. **Row 2**: The GTV segmentations show good overlap between the DL Network and both ROs. **Row 3**: The DL Network overestimated the extent of the sarcoma by adding muscular tisse to the GTV. Moreover, it falsely predicted sarcoma in a small isle within the bone.

based machine learning, rather than GTVs directly suitable for the RT planning process. Our agreement maps revealed that the ROs, who used this latter approach for VOI definition, deliberately included equivocal areas in surrounding tissues for safe treatment planning. Moreover, the results confirmed decreased reproducibility between the GTs and the GTV definitions of the ROs. Here, the median DSC was 0.82 (IQR 0.10) for the manual segmentations.

Even though the algorithm was not trained to generate GTVs for direct application in RT, the DL-based VOI predictions achieved clinical feasibility in a subset of patients. Thus, 7/20 (35%) and 4/20 (20%) segmentations in the benchmark cohort were deemed clinically suitable by the two ROs. A case-by-case analysis revealed that these cases showed sharp VOI definitions of the gross tumor without suspected infiltrative behavior. Overall, these observations demonstrate the need for standardized definitions of the segmentation directed towards the intended use of the future model. Therefore, we believe that our model can guide future AI application for feature extraction and bounding box definition of DL models. However, specialized clinical application would mandate a specific retraining of the presented algorithm.

Importantly, clinical applications of accurate and objective VOI delineation could also be impactful for diagnostic radiology. There is a critical need to develop more reliable forms of volumetric tumor characterization and response assessment models specific for STS, as these irregularly shaped tumors are often poorly characterizable using RECIST 1.1 criteria [36]. Future directions may allow the use of machine learning-based radiomic signatures to inform on more specific STS response assessment for treatment effect. The analysis of the VOI-related signatures may allow better differentiation of pseudoprogression with increased treatment effect from true progression [37].

While the presented model showed high segmentation agreement with GTs (median DSC values greater than 0.8), time assessment within the benchmark study revealed that DL-assisted VOI definition did not yield time savings. Interviews of the two ERs revealed that manual review and adjustment on a slice-by-slice basis was time consuming, especially for large or complex and irregular target volumes. Thus, time-related advantages of using DLBAS may necessitate direct utilization of the model predictions for radiomic feature extraction without prior manual correction.

Our study bears several limitations: Our cohort comprised a diverse set of STS subhistologies. The generation of sufficiently large patient sets remains challenging due to STS rareness and subhistology diversity. Subhistology-specific segmentation models may provide increased reproducibility as subhistologies can result in distinct sequence-dependent MRI signatures, e.g. myxoid liposarcomas appear T2 hyperintense. Thus, future studies need to address the development of subistology-specific models with sufficiently large datasets. Also, to achieve effective neural network training, we only included STS located extraperitoneally or outside of the subperitoneal compartment. Thus the datasets comprised mostly tumors located in the upper and lower extremities. A further 14,4% of tumors across both datasets were located in the muscular and subcutaneous tissues of the pelvis, and deemed fit for inclusion as part of the musculoskeletal system of the extremities. As such, the presented model was not designed to predict visceral STS located in the retroperitoneal space.

A further limitation of the study design is the use of only a single fat-saturated contrast-enhanced T1-weighted (T1-CE) MRI sequence. The MRI sequence parameters are shown in **Table S7**. In a clinical setting, GTV segmentations for RT treatment planning require a dedicated planning CT as well as further MRI sequences for the best possible unambiguous tumor volume delineation. The presented segmentation algorithm was primarily designed to reproducibly predict the primary gross tumor as VOI for Radiomics analyses, as has been the focus of several recent MRI-based studies, such as for meningioma [38] or hypopharyngeal cancer [39]. While this work also evaluated the usefulness of DLBAS predictions for direct clinical application in the clinical benchmark substudy, the decision for using only a single T1-CE MRI sequence was in accordance with contemporary guidelines for GTV definition [4,7]. This approach yielded the largest available sample size

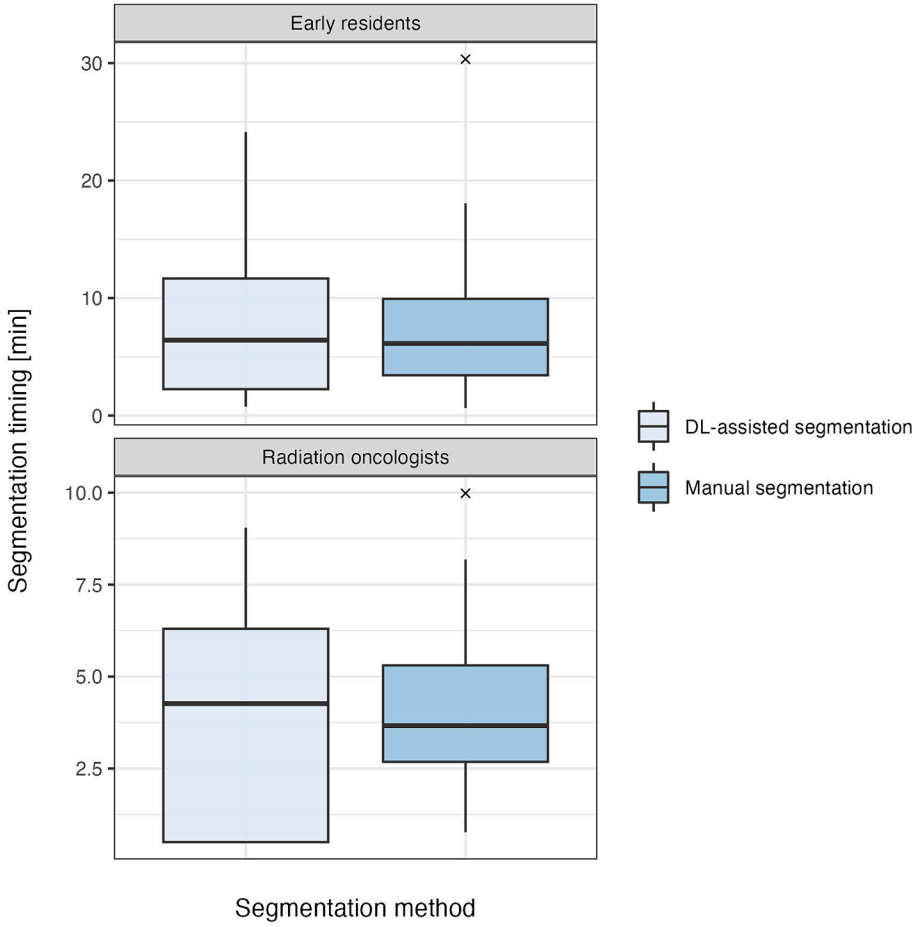

Fig. 5. Boxplot of segmentation timing comparison.

for our training and testing cohort, which is necessary to develop a robust DLBAS model. Nevertheless, future studies focusing on clinical GTV definition should include multiple MRI sequences.

The design of the segmentation benchmark inherited several confounding factors. To minimize case-specific learning effects from repetitive segmentations, a 4-week interval was chosen alongside a crossover design. There remains the risk that each second segmentation task was performed faster than the first. When comparing segmentation timings for each respective imaging study, we found no statistically significant difference in the segmentation duration. Finally, our model requires the upfront definition of a bounding box, thus yielding a semi-automatic procedure. For completely automatic predictions, a localization step is necessary.

In conclusion, we provide a U-Net-based segmentation algorithm for MRI-based volume of interest prediction of extremity soft-tissue sarcomas. The resulting gross tumor delineations can be used as a reproducible basis for radiomic feature extraction or for bounding box definitions for DL analyses. Despite the observed values for segmentation congruence between algorithm predictions and clinically intended delineations within the benchmark study, a specific retraining and addition of further MRI sequences is necessary to generate expert GTV definitions directly applicable for the RT planning process. The segmentation algorithm can be accessed at https://github.com/RadOnc-AI/SoftTissueSarcomaSegmentation.

Ethics approval and content to participate

Ethics approval was granted by the ethics committee of the Technical University of Munich (466/6s).

Dataset availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request. The segmentation algorithm can be accessed at https://github.com/RadOnc-AI/SoftTissueSarcomaSegmentation.

Funding

This work was funded by the Wilhelm Sander Stiftung (2022.032.1). Open Access funding enabled and organized by DEAL agreement.

CRediT authorship contribution statement

Jan C. Peeken: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lucas Etzel:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tim Tomov:** Writing – review & editing, Software, Investigation, Data curation. **Stefan Münch:** Writing – review & editing, Investigation, Data curation. **Lars Schüttrumpf:** Writing – review & editing, Investigation, Data curation. **Julius H. Shaktour:** Writing – review & editing, Investigation, Data curation. **Johannes Kiechle:** Writing – review & editing, Visualization, Investigation, Formal analysis. **Carolin Knebel:** Writing – review & editing, Investigation. **Stephanie K. Schaub:** Writing – review & editing, Investigation. **Nina A. Mayr:** Writing – review & editing, Investigation. **Henry C. Woodruff:** Writing – review & editing, Investigation. **Philippe Lambin:** Writing – review & editing, Investigation. **Alexandra**

S. Gersing: Writing – review & editing, Investigation. **Denise Bernhardt:** Writing – review & editing, Investigation. **Matthew J. Nyflot:** Writing – review & editing, Investigation. **Bjoern Menze:** Writing – review & editing, Investigation. **Stephanie E. Combs:** Writing – review & editing, Supervision, Project administration. **Fernando Navarro:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2024.110338.

References

[1] Rosenberg SA, Tepper J, Glatstein E, Costa J, Baker A, Brennan M, et al. The treatment of soft-tissue sarcomas of the extremities: prospective randomized evaluations of (1) limb-sparing surgery plus radiation therapy compared with amputation and (2) the role of adjuvant chemotherapy. Ann Surg 1982;196:305–15.

[2] Yang JC, Chang AE, Baker AR, Sindelar WF, Danforth DN, Topalian SL, et al. Randomized prospective study of the benefit of adjuvant radiation therapy in the treatment of soft tissue sarcomas of the extremity. J Clin Oncol 1998;16:197–203.

[3] O'Sullivan B, Davis AM, Turcotte R, Bell R, Catton C, Chabot P, et al. Preoperative versus postoperative radiotherapy in soft-tissue sarcoma of the limbs: a randomised trial. Available from The Lancet [Internet] 2002;359:2235–41. https://linkinghub.elsevier.com/retrieve/pii/S0140673602092929.

[4] Salerno KE, Alektiar KM, Baldini EH, Bedi M, Bishop AJ, Bradfield L, et al. Radiation therapy for treatment of soft tissue sarcoma in adults: executive summary of an ASTRO clinical practice guideline. Pract Radiat Oncol 2021;11:339–51.

[5] Haas RLM, Delaney TF, O'Sullivan B, Keus RB, Pechoux CL, Olmi P, et al. Radiotherapy for management of extremity soft tissue sarcomas: why, when, and where? Int J Radiat Oncol Biol Phys 2012;84:572–80.

[6] Roberge D, Skamene T, Turcotte RE, Powell T, Saran N, Freeman C. Inter- and intra-observer variation in soft-tissue sarcoma target definition. Cancer/Radiotherapie 2011;15:421–5.

[7] Wang D, Bosch W, Roberge D, Finkelstein SE, Petersen I, Haddock M, et al. RTOG sarcoma radiation oncologists reach consensus on gross tumor volume and clinical target volume on computed tomographic images for preoperative radiotherapy of primary soft tissue sarcoma of extremity in radiation therapy oncology group studies. Int J Radiat Oncol Biol Phys [Internet]. 2011;81:e525–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0360301611005591.

[8] Peeken JC, Wiestler B, Combs SE. Image-guided radiooncology: the potential of radiomics in clinical application. 2020.

[9] Peeken JC, Bernhofer M, Spraker MB, Pfeiffer D, Devecka M, Thamer A, et al. CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. Radiother Oncol [Internet]. 2019;135:187–96. Available from: https://doi.org/10.1016/j.radonc.2019.01.004 https://linkinghub.elsevier.com/retrieve/pii/S0167814019300088.

[10] Peeken JC, Spraker MB, Knebel C, Dapper H, Pfeiffer D, Devecka M, et al. Tumor grading of soft tissue sarcomas using MRI-based radiomics. EBioMedicine 2019;48:332–40.

[11] Peeken JC, Asadpour R, Specht K, Chen EY, Klymenko O, Akinkuoroye V, et al. MRI-based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy. Radiother Oncol 2021;164:73–82.

[12] Peeken JC, Neumann J, Asadpour R, Leonhardt Y, Moreira JR, Hippe DS, et al. Prognostic assessment in high-grade soft-tissue sarcoma patients: a comparison of semantic image analysis and radiomics. Cancers 2021;13.

[13] Navarro F, Dapper H, Asadpour R, Knebel C, Spraker MB, Schwarze V, et al. Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using mr imaging. Cancers 2021;13:1–14.

[14] Poirot MG, Caan MWA, Ruhe HG, Bjørnerud A, Groote I, Reneman L, et al. Robustness of radiomics to variations in segmentation methods in multimodal brain MRI. Sci Rep 2022;12:16712.

[15] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. 2015; Available from: http://arxiv.org/abs/1505.04597.

[16] Roy AG, Navab N, Wachinger C. Concurrent spatial and channel squeeze & excitation in fully convolutional networks. 2018; Available from: http://arxiv.org/abs/1803.02579.

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition. IEEE Computer Society; 2016. p. 770–8.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017; Available from: http://arxiv.org/abs/1706.03762.

[19] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. 2019; Available from: http://arxiv.org/abs/1912.01703.

[20] Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approachin pre-clinical, clinical and laboratory studies. Biochem Med 2021;31:1–27.

[21] The plastimatch project ("plastimatch") maintained by the general hospital corporation inc. ("MGH"). [Internet]. Available from: https://plastimatch.org/.

[22] R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/.

[23] RStudio Team. RStudio: integrated development environment for r [Internet]. Boston, MA: RStudio, PBC; 2020. Available from: http://www.rstudio.com/.

[24] Xie Y, Dervieux C, Riederer E. R markdown cookbook [Internet]. Boca Raton, Florida: Chapman; Hall/CRC; 2020. Available from: https://bookdown.org/yihui/rmarkdown-cookbook.

[25] Xie Y. Bookdown: authoring books and technical documents with r markdown [Internet]. 2023. Available from: https://CRAN.R-project.org/package=bookdown.

[26] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. J Open Sour Softw 2019;4:1686.

[27] Kassambara A. Ggpubr: ggplot2 based publication ready plots [Internet]. 2023. Available from: https://rpkgs.datanovia.com/ggpubr/.

[28] Kassambara A. Rstatix: pipe-friendly framework for basic statistical tests [Internet]. 2023. Available from: https://rpkgs.datanovia.com/rstatix/.

[29] Gamer M, Lemon J, <puspendra.pusp22@gmail.com> IFPS. Irr: various coefficients of interrater reliability and agreement [Internet]. 2019. Available from: https://www.r-project.org.

[30] Yang F, Simpson G, Young L, Ford J, Dogan N, Wang L. Impact of contouring variability on oncological PET radiomics features in the lung. Sci Rep 2020;10.

[31] Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol 2018;57:1070–4.

[32] Huang Q, Lu L, Dercle L, Lichtenstein P, Li Y, Yin Q, et al. Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. J Med Imaging 2017;5:1.

[33] Fontaine P, Andrearczyk V, Oreiller V, Abler D, Castelli J, Acosta O, et al. Cleaning radiotherapy contours for radiomics studies, is it worth it? A head and neck cancer study. Clin Transl Radiat Oncol 2022;33:153–8.

[34] Rizzetto F, Calderoni F, Mattia CD, Defeudis A, Giannini V, Mazzetti S, et al. Impact of inter-reader contouring variability on textural radiomics of colorectal liver metastases. Eur Radiol Exper 2020;4.

[35] Fontaine P, Andrearczyk V, Oreiller V, Castelli J, Jreige M, Prior JO, et al. Fully automatic head and neck cancer prognosis prediction in PET/CT. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer Science; Business Media Deutschland GmbH; 2021. p. 59–68.

[36] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009 Jan;45:228–47.

[37] Messiou C, Bonvalot S, Gronchi A, Vanel D, Meyer M, Robinson P, et al. Evaluation of response after pre-operative radiotherapy in soft tissue sarcomas; the european organisation for research and treatment of cancer - soft tissue and bone sarcoma group (EORTC - STBSG) and imaging group recommendations for radiological examination and reporting with an emphasis on magnetic resonance imaging. Eur J Cancer 2016;56:37–44.

[38] Lin YC, Lin G, Pandey S, Yeh CH, Wang JJ, Lin CY, et al. Fully automated segmentation and radiomics feature extraction of hypopharyngeal cancer on MRI using deep learning. Eur Radiol 2023;33:6548–56.

[39] Yang L, Wang T, Zhang J, Kang S, Xu S, Wang K. Deep learning–based automatic segmentation of meningioma from T1-weighted contrast-enhanced MRI for preoperative meningioma differentiation using radiomic features. BMC Med Imaging 2024;24.

[40] Tixier F, Rest CCL, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52:369–78.

[41] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 2020;295:328–38.