

RectoMap: a generalizable deep learning tool for rectal cancer and mesorectum MRI segmentation

Simone Perra

University of Padua

Filippo Crimi

University of Padua

Valentina Visani

University of Padua

Niccolò Sion

University of Padua

Matteo Prezioso

University of Padua

Filippo De Lucchi

University of Padua

Sussan Giovanni

University of Padua

Francesco Celotto

University of Padua

Claudio Coco

Università Cattolica del Sacro Cuore

Giuditta Chiloiro

Agostino Gemelli University Polyclinic

Marco Scarpa

University of Padua

Emilio Quaia

University of Padua

Mauro Mattace Raso

Istituto Nazionale Tumori IRCCS "Fondazione G. Pascale"

Annalisa Marteddu

University of Cagliari

Luca Saba

University of Cagliari

Daniela Rega

Istituto Nazionale Tumori IRCCS "Fondazione G. Pascale"

Angelo Restivo

University of Cagliari

Simona Deidda

University of Cagliari

Gaya Spolverato

University of Padua

Marco Castellaro

marco.castellaro@unipd.it

University of Padua

Research Article

Keywords: Rectal Neoplasms, Magnetic Resonance Imaging, Deep Learning, Image Processing, Computer-Assisted, Multicenter Studies as Topic, Reproducibility of Results

Posted Date: July 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10432717/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Title

RectoMap: a generalizable deep learning tool for rectal cancer and mesorectum MRI segmentation

Abstract

Purpose: To develop and evaluate RectoMap, a robust open-source deep learning pipeline for fully automatic 3D segmentation of rectal cancer and mesorectum on T2-weighted MRI, with a specific focus on generalization across MRI scanner vendors in heterogeneous multi-institutional data.

Methods: In this retrospective multi-institutional study, 226 patients with locally advanced rectal cancer were imaged across 19 scanner models from four vendors (Siemens, Philips, GE, Hitachi) at 1.5T and 3T. Two complementary 3D architectures (nnUNet and U-MambaBot) were trained with 5-fold cross-validation under standard and extended MRI-specific data augmentation. Four ensemble strategies based on softmax averaging and the STAPLE algorithm were evaluated on an in-distribution (ID; $n = 45$) and an out-of-distribution (OOD; $n = 63$, previously unseen vendors) test set. Segmentation performance was quantified using the Dice similarity coefficient (DSC) and the 95th-percentile Hausdorff distance.

Results: 226 patients (mean age, 62.7 years $\pm$ 12.7 [SD], 139 men) were analyzed. In the ID setting, two-level STAPLE aggregation with extended augmentation achieved the highest rectal cancer DSC (0.767 ± 0.119), while full-model aggregation yielded the best mesorectum DSC (0.794 ± 0.131). MRI-specific augmentation consistently improved mesorectum segmentation. In the OOD setting, despite vendor shift, full-model aggregation achieved DSC 0.798 ± 0.126 for rectal cancer and 0.776 ± 0.161 for mesorectum, demonstrating stable cross-vendor performance.

Conclusion: RectoMap enables accurate and robust automatic segmentation of rectal cancer and mesorectum across heterogeneous MRI data. It is publicly released as an open-source tool that can be applied directly or fine-tuned on local datasets, providing a reliable baseline even when limited institutional data are available.

Keywords: *Rectal Neoplasms, Magnetic Resonance Imaging, Deep Learning, Image Processing, Computer-Assisted, Multicenter Studies as Topic, Reproducibility of Results*

1. Introduction

Locally advanced rectal cancer, a subtype of colorectal cancer that ranks as the third most frequently diagnosed cancer and the second leading cause of cancer-related mortality globally (1), typically requires multimodal management, including neoadjuvant chemoradiotherapy followed by total mesorectal excision (TME) (2). Within this clinical workflow, magnetic resonance imaging (MRI) plays a central role, providing high soft-tissue

contrast and enabling detailed visualization of both the primary tumor and the mesorectal compartment, which are critical for accurate staging, treatment planning and response assessment (3-6).

Accurate segmentation of the rectal tumor and mesorectum is clinically relevant for multiple reasons. Tumor delineation enables quantitative estimation of tumor volume and tumor regression grade after chemoradiotherapy, both recognized as key prognostic biomarkers that guide individualized therapeutic decisions such as “watch-and-wait” strategies for complete responders or treatment intensification for non-responders (7). The tumor contour also defines the gross tumor volume used in radiotherapy planning, ensuring accurate target coverage while minimizing dose exposure to adjacent organs at risk (8).

Complementarily, segmentation of the mesorectum, in the surgical setting, supports preoperative assessment of resectability and aids in planning complete TME while preserving critical structures (9). Furthermore, emerging radiomic studies have demonstrated that mesorectal texture and shape features capture biologically relevant information about the peritumoral microenvironment, correlating with treatment response and outcomes (10-13).

Despite these benefits, manual segmentation of the tumor and mesorectum remains time-consuming and operator-dependent, often characterized by considerable inter-observer variability (14).

In recent years, deep learning-based approaches have shown substantial promise in overcoming these challenges, offering automated and reproducible solutions for medical image segmentation (15-18).

However, their clinical applicability remains constrained by significant challenges, particularly in addressing the variability inherent in MRI scans. Factors such as scanner bias field, sequence parameters, acquisition protocols, and manufacturer differences, collectively referred to as manufacturer shift, introduce substantial heterogeneity that can lead to degraded model performance when applied to unseen images acquired in diverse settings, a challenge known as out-of-distribution (OOD) generalization.

Previous studies on MRI-based rectal cancer segmentation have reported promising but highly variable results, largely influenced by dataset size, scanner heterogeneity, and evaluation design.

Works conducted on single-scanner datasets generally achieved the highest accuracy, with mean Dice similarity coefficient (DSC) values of 0.858 (43 subjects) (19), 0.74 (93 subjects) (20), 0.74 (43 subjects) (21), and a median DSC of 0.670 (81 subjects) (22).

When moving to multi-scanner datasets, performance tends to decrease. Studies including two to four scanners reported median DSC values of 0.70 (140 subjects) (23) and 0.727 (201 subjects) (24), whereas another three-scanner dataset reported a mean DSC of 0.856 (65 subjects) (25).

However, these results derive from homogeneous, single- or few-scanner cohorts, limiting generalizability to real-world multi-vendor settings.

Literature evidence underscores the importance of addressing such challenges. For instance, a study testing manufacturer shift on cardiac left ventricle segmentation observed a significant decline in DSC performance, from 0.906 to 0.595 (26).

In the context of rectal cancer segmentation, studies explicitly assessing manufacturer shift have shown marked performance degradation when testing on scanners unseen during training. Reported drops include a median DSC decline from 0.77 to 0.59 (192 subjects) (27) and from 0.73 to 0.69 (100 subjects) (28). A recent multi-center (14 institutions) involving 1212 subjects reported median DSC values of 0.68 (for the in-distribution test set) and 0.66 (for the out-of-distribution test set) (29), despite its substantially larger cohort.

For mesorectum segmentation, available evidence remains limited. The largest published study reported a median DSC of 0.917 (201 subjects) (24), although without evaluation under out-of-distribution conditions.

These findings underscore the difficulty of generalizing segmentation models to heterogeneous, multi-scanner clinical data.

To address these limitations, we developed and publicly released RectoMap, a novel open-source tool for automatic segmentation of rectal tumor and mesorectum on 3D T2-weighted MRI (the tool is openly available on GitLab at [https://\[institutional GitLab URL masked for double-anonymous review; will be restored upon acceptance\]](https://[institutional GitLab URL masked for double-anonymous review; will be restored upon acceptance])). Rather than a benchmarking exercise, RectoMap is intended as a generalizable clinical tool: the primary contribution of this work is a cross-vendor-robust, ready-to-use segmentation pipeline, with the underlying ensembling and augmentation comparisons serving to justify and support this design choice.

2. Materials and methods

Dataset

This retrospective study was approved by the local Ethics Committee at each participating institution and conducted in accordance with the Declaration of Helsinki, with written informed consent obtained from all patients. This study included MRI scans from 226 patients (mean age $\pm$ SD: 62.7 $\pm$ 12.7 years, M/F: 139/87) with histologically confirmed LARC. All patients underwent MRI examination as part of their baseline diagnostic evaluation prior to neoadjuvant therapy. The inclusion criteria were as follows: patients aged over 18 years, histologically proven locally advanced rectal adenocarcinoma located within 12 cm from the anal verge and candidates for neoadjuvant treatments who were able to undergo TME surgery. Data were retrospectively collected from four institutions, representing a geographically and technically heterogeneous cohort:

- *Institution A* - 106 subjects (coordinating center);
- *Institution B* - 51 subjects;
- *Institution C* - 36 subjects;
- *Institution D* - 33 subjects.

For each institution, the study protocol was approved by the respective local Ethics Committee and all procedures were conducted in accordance with the Declaration of Helsinki. All patients provided written informed consent for the use of anonymized imaging data for research purposes. All DICOM data were anonymized at the site of origin prior to transfer and analysis.

MRI acquisition and annotation

For all subjects, baseline axial T2-weighted MRI sequences were acquired prior to the start of neoadjuvant therapy. Acquisitions were performed across 19 distinct MRI scanner models, spanning four major vendors (Siemens, Philips, GE and Hitachi), with 183 examinations acquired at 1.5T and 43 at 3T. This diversity resulted in marked heterogeneity in acquisition protocols. An overview of the available acquisition parameters (including TR, TE, flip angle, number of averages, and coil configuration) for each contributing institution is provided in [Table S1](#) in the *Supplementary* section.

Manual annotations of the rectal tumor and mesorectum were performed by a team of experienced radiologists using 3D Slicer (30), an open-source software platform for medical

image processing. These contours served as the ground truth labels for model training and evaluation. Examples of manual delineations are illustrated in [Figure 1](#). Each case was segmented by a single radiologist; discordant or uncertain delineations were subsequently reviewed and resolved by consensus adjudication with a senior radiologist to establish the final reference contour.

Dataset heterogeneity

The dataset exhibits substantial heterogeneity, encompassing differences not only in MRI acquisition parameters but also in anatomical presentation and preparation protocols. Scans were obtained across multiple institutions using various MRI vendors, models, and field strengths, resulting in broad variability in image quality, contrast, and spatial resolution. In particular, the in-plane voxel spacing ranged from $(0.18 \times 0.18) \text{ mm}^2$ to $(1.25 \times 1.25) \text{ mm}^2$, while the slice thickness varied between 3 mm and 8 mm. The field of view extended approximately from $(166 \times 200) \text{ mm}^2$ to $(461 \times 461) \text{ mm}^2$ in-plane, and from 78 mm to 404 mm along the z-axis, depending on the scanner configuration and patient anatomy. These parameters are summarized in [Table 1](#), which reports representative ranges of the principal images characteristics across all scanners.

Beyond acquisition variability, the anatomical and structural heterogeneity of the segmented regions further increases the dataset's complexity. The rectal tumor demonstrates wide variations in size, morphology, and localization, from small, annular lesions to bulky masses, and can involve different rectal wall layers or extend beyond the muscularis propria.

In some examinations, the rectum was distended with ultrasound gel, a practice commonly adopted in rectal MRI protocols to improve luminal contrast and facilitate precise visualization of the rectal wall and intraluminal tumor components (31). In other cases, no gel was administered, resulting in partially collapsed lumens or irregular wall apposition. This variability in rectal distension influences overall image appearance, alters the spatial relationship between the tumor and the mesorectal fascia, and introduces subtle intensity and shape variations that pose additional challenges for automatic segmentation algorithms. Such differences mirror routine clinical practice, where rectal filling protocols are not standardized across institutions or vendors, and thus provide a realistic testing ground for developing robust, generalizable deep learning models.

Representative examples of these variations, including differences in tumor morphology, image contrast, and rectal filling conditions, are illustrated in [Figure 1](#), which highlights the wide spectrum of appearances encountered across scanners and acquisition protocols.

Experimental setups

To comprehensively assess model robustness and generalization, two complementary experimental setups were designed: an in-distribution (ID) and an out-of-distribution (OOD) configuration.

These setups were conceived to evaluate performance both under conditions reflecting known scanner domains and under those involving previously unseen MRI vendors, thus providing a controlled framework for testing in-domain consistency and cross-scanner generalization.

- **ID setting:** in the ID configuration, the dataset was randomly split into 181 subjects for training and 45 subjects for testing, ensuring that all MRI scanners represented in the test set were also present in the training data. This setup simulates a scenario where the model encounters scanner domains similar to those seen during training.

The distribution of scanner models across the training and testing subsets is detailed in [Table 2](#).

- **OOD setting:** the OOD setup was specifically designed to test the model’s ability to generalize to previously unseen scanner types. In this configuration, models were trained on 163 subjects acquired from Siemens (n = 109), Toshiba (n = 31), and Philips (n = 23) scanners, and tested on a completely unseen set of 63 subjects acquired using GE (n = 62) and Hitachi (n = 1) systems. [Table 2](#) reports the full distribution of scanner models and subjects included in the OOD training and testing partitions.

Tested Architectures

To ensure a fair and representative evaluation of current state-of-the-art architectures for volumetric MRI segmentation, we selected two complementary 3D deep learning models that differ in design philosophy and feature representation capacity. Specifically, we employed nnUNet (16), a self-configuring CNN-based framework, and U-MambaBot (17), a hybrid architecture integrating convolutional and state-space modeling for improved long-range context representation. A detailed summary of the training configurations for both architectures (including batch size, patch size, learning rate scheduler, initial learning rate, loss function, optimizer, and weight decay) is provided in [Supplementary Table S2](#).

Ensembling Strategies

To improve segmentation robustness and leverage the complementary strengths of different network architectures, four ensembling configurations were designed.

Both nnUNet and U-MambaBot were independently trained using identical 5-fold cross-validation splits. Predictions from the two architectures were subsequently combined using different fusion strategies based on averaging and/or the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (32), which estimates a probabilistic consensus segmentation from multiple inputs.

Let:

- $P_{a,f}$ = prediction of architecture a (nnUNet or U-MambabBot) for fold f
- $avg(\cdot)$ = voxel-wise averaging of softmax probability maps across folds
- $STAPLE(\cdot)$ = fusion using the STAPLE algorithm

Then, the four ensemble configurations are defined as follows:

- **Configuration 1 - Architecture-wise averaging:** for each architecture, the softmax predictions $\tilde{P}$ from its five folds were averaged and then combined using argmax producing one mask for nnUNet and one for U-MambaBot). These two architecture-level results were then merged using STAPLE to obtain the final segmentation:

$$S_{final} = STAPLE\left((avg(\tilde{P}_{nnUNet,1...5})), (avg(\tilde{P}_{U-MambaBot,1...5}))\right)$$

- **Configuration 2 - Fold-wise best model selection:** for each cross-validation fold, the model (either nnU-Net or U-MambaBot) achieving the highest validation Dice coefficient was selected. Test predictions from these five “best-performing” models were then combined using STAPLE:

$$S_{final} = STAPLE(P_{Best,1...5})$$

- **Configuration 3 - Two-level STAPLE aggregation:** for each architecture, the five fold predictions were first combined using STAPLE, resulting in one consolidated segmentation per network. These two architecture-level outputs were then ensembled again via STAPLE:

$$S_{final} = STAPLE\left(STAPLE(P_{nnUNet,1...5}), STAPLE(P_{U-MambaBot,1...5})\right)$$

- **Configuration 4 - Full-model aggregation:** predictions from all ten models (five folds $\times$ two architectures) were combined directly through a single STAPLE step. This approach maximizes model diversity while bypassing intermediate averaging:

$$S_{final} = STAPLE(P_{nnUNet,1...5}, P_{U-MambaBot,1...5})$$

Data augmentation

To enhance model generalization and robustness against scanner- and acquisition-related variability, two levels of data augmentation were applied during training.

All experiments included the default nnUNet augmentation pipeline, hereafter referred to as *base augmentation*, while a subset of experiments incorporated an *extended augmentation* scheme designed to simulate MRI-specific artifacts. Both augmentation configurations were applied independently within each training fold and used identical random seeds for reproducibility:

- *base augmentation* (default nnUNet data augmentation): the base augmentation included the standard geometric and intensity transformations provided by nnUNet, which are known to improve model invariance to orientation, scale, and appearance variability. See [Table 3](#) for more details.
- *extended MRI-specific augmentation*: in the extended augmentation setting, the base augmentations were combined with additional MRI-specific artifact simulations implemented via TorchIO (33) and Global Intensity Non-linear (GIN) augmentation-based transformations (34). These augmentations were designed to emulate real-world acquisition imperfections. See [Table 3](#) for more details.

Application across experimental settings

Both ID and OOD experiments were trained and evaluated twice:

- with base augmentation only, using the standard nnUNet data augmentation pipeline;
- with extended augmentation, combining base and MRI-specific artifact simulations.

In total, each of the four ensembling configurations was evaluated under both augmentation regimes, resulting in a comprehensive performance comparison across 8 experimental configurations per setting.

Evaluation metrics

For model evaluation, two metrics were used: Dice similarity coefficient (DSC) and the 95th percentile Hausdorff distance (HD95).

The DSC evaluates the overlap between the predicted and ground truth segmentation mask, ranging from 0 (no overlap) to 1 (perfect overlap). Its formula is:

$$DSC = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

where A is the set of predicted pixels, B is the set of ground truth pixels, and |A| and |B| are the total number of pixels in each set.

The HD measures the greatest distance between the boundaries of predicted and ground truth segmentations. The 95th percentile HD (HD95) reduces the impact of outliers by focusing on the largest distances within the 95th percentile. Its formula is:

$$HD(A, B) = \max(\max_{a \in A} \min_{b \in B} \|a - b\|, \max_{b \in B} \min_{a \in A} \|a - b\|)$$

where A and B are the predicted and ground truth boundaries, and $\|a - b\|$ is the Euclidean distance.

3. Results

Two-hundred and twenty-six (226) patients (mean age, 62.7 years $\pm$ 12.7 [SD], 139 men) were analyzed. [Tables 4](#) and [5](#) summarize the quantitative performance of the four ensembling configurations in both ID and OOD settings, with and without the extended MRI-specific data augmentation.

In the ID experiment, the two-level STAPLE aggregation ensemble configuration (**configuration 3**) with extended data augmentation achieved the best mean DSC for rectal cancer (0.767 ± 0.119), whereas the full-model aggregation (**configuration 4**) configuration yielded the highest DSC for the mesorectum (0.794 ± 0.131).

In the OOD setting, despite the domain shift induced by unseen scanner vendors, performance remained stable. The full-model aggregation configuration (**configuration 4**) without extended data augmentation achieved the best DSC for rectal cancer (0.798 ± 0.126), while the same configuration with augmentation provided the best performance for the mesorectum (0.776 ± 0.161). [Figure 2](#) provides a qualitative example of model predictions using both ID and OOD configurations.

For the fold-wise model selection (ensembling configuration 2), the model with the highest validation DSC, computed from the best-performing checkpoint within each fold, was selected between nnUNet and U-MambaBot. To provide transparency regarding the influence of architecture choice on the final ensemble, [Tables S3](#) (ID) and [S4](#) (OOD) in the *Supplementary* section report the validation DSC obtained by each architecture across the five folds. For each fold, the selected architecture is highlighted as it directly contributed to the fold-wise ensemble prediction.

Qualitative review of the lowest-performing cases indicated that segmentation errors were concentrated in subjects with small or annular rectal tumors, collapsed or under-distended rectal lumens (in the absence of ultrasound gel filling), and examinations acquired on scanner models markedly underrepresented in the training data. These findings suggest that tumor size, luminal distension, and scanner representation are the principal drivers of residual segmentation failures.

4. Discussions

This study introduces *RectoMap*, a robust and publicly available deep learning tool for the automatic segmentation of rectal tumor and mesorectum on 3D T2-weighted MRI. To determine the optimal design on which to build the final *RectoMap* model, we conducted a systematic evaluation of multiple network ensembling configurations and MRI-specific data augmentation strategies across a uniquely heterogeneous multi-scanner dataset. By analysing how these elements interact and contribute to robustness, we identified the combination that delivers the most reliable performance both ID and under OOD scanner shifts.

Impact of data augmentation

Across the ID experiments, the introduction of MRI-specific extended data augmentation consistently improved segmentation performance for both target structures, enhancing DSC accuracy and producing smoother, anatomically plausible boundaries. The gains were particularly evident for the mesorectum. This confirms that simulating acquisition imperfections (such as bias field inhomogeneity, motion artifacts, ghosting, and intensity nonlinearities) helps the models develop a more robust feature representation. These findings align with prior evidence indicating that MRI-specific augmentation improves robustness in heterogeneous multi-scanner datasets (35-36).

Generalization to unseen vendors

In the OOD setting, rectal tumor segmentation achieved a higher DSC compared to the ID configuration. This behavior can be partly attributed to anatomical differences between the two test cohorts. As illustrated in [Figure S5 \(Supplementary section\)](#), tumors in the OOD set had smaller average volumes, which typically correspond to simpler shapes, making them intrinsically easier to segment and likely contributing to the improved performance.

For the mesorectum, however, the opposite trend was observed: mesorectal volumes were larger in the OOD cohort, increasing shape variability. This anatomical difference helps explain why mesorectum performance was higher in the ID setting, where the structures were generally smaller and more regular.

In addition to anatomical factors, scanner-related characteristics also played a role. The OOD test set, although composed of previously unseen vendors, exhibited relatively low intra-vendor variability, with the majority of scans acquired on a GE model. This homogeneity may have attenuated the performance degradation typically associated with vendor shift. Additionally, because most training data were acquired on Siemens scanners, the models may have developed more stable and conservative feature representations, (effectively acting as a form of implicit regularization) which can support generalization when tested on more uniform data distributions.

Taken together, these findings suggest that the observed OOD performance is influenced not only by scanner-related factors but also by anatomical and structural properties of the target regions.

Role of ensemble design

Among the four ensemble configurations tested, the **two most effective** across both ID and OOD settings were:

- **Configuration 3:** this strategy first consolidates the predictions within each architecture and then fuses the resulting architecture-level segmentations. By doing so, it effectively captures the characteristic strengths of each network while suppressing fold-specific noise.
- **Configuration 4:** this configuration aggregates all fold predictions from both architectures in a single STAPLE step, maximizing model diversity. This broader diversity appears particularly advantageous under OOD conditions.

Choice of the final RectoMap model

Based on the above findings, *RectoMap* is built upon Configuration 3 (two-level STAPLE aggregation) combined with the extended MRI-specific augmentation. The final model is trained on the complete in-distribution training set.

To maximize utility for clinical and research applications, we additionally retrained the final model on the entire dataset (all 226 subjects, including previous test subjects) to enlarge the training distribution. The resulting weights are publicly released, and users can:

- download the model checkpoints;
- perform inference directly through the RectoMap pipeline;
- fine-tune the network on local institutional datasets to adapt to specific acquisition protocols.

We provide an open-source implementation on GitLab ([https://\[institutional GitLab URL masked for double-anonymous review; will be restored upon acceptance\]](https://[institutional GitLab URL masked for double-anonymous review; will be restored upon acceptance])).

Inference time for a full 3D T2w MRI volume is approximately 82 seconds on an NVIDIA GeForce RTX 3090, enabling practical integration into clinical workflows. We include in [Table S6 \(Supplementary section\)](#) a detailed summary of the training characteristics of the best-performing configuration on which RectoMap is based.

Conclusions

RectoMap represents an important step toward robust, generalizable automated segmentation tools for rectal cancer. Future directions include expanding the dataset to include more centers and vendors, prospective validation in clinical workflows, adaptation to additional MRI sequences, computational optimization for real-time deployment, and integration of continuous learning or federated learning approaches to enhance generalization and preserve data privacy. These developments could facilitate broader clinical adoption and potentially improve therapeutic decision-making and planning in rectal cancer management. Despite competitive performance, limitations should be acknowledged. The dataset, although multi-scanner and heterogeneous, remains relatively skewed toward a predominant vendor (Siemens). The impact of *RectoMap* on treatment planning, surgical decision-making, or radiotherapy workflows has not yet been directly assessed, limiting insight into its clinical utility and potential benefits for patient outcomes.

References

1. Sung H, Ferlay J, Siegel RL et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 71:209-249 <https://doi.org/10.3322/caac.21660>
2. Troester AM, Gaertner WB (2024) Contemporary management of rectal cancer. *Surg Open Sci* 18:17-22 <https://doi.org/10.1016/j.sopen.2024.01.009>
3. Beets-Tan RGH, Beets GL (2004) Rectal cancer: Review with emphasis on MR imaging. *Radiology* 232:335-346 <https://doi.org/10.1148/radiol.2322021326>
4. Kaur H, Choi H, You YN et al (2012) MR imaging for preoperative evaluation of primary rectal cancer: Practical considerations. *Radiographics* 32:389-409 <https://doi.org/10.1148/rg.322115122>
5. Lambregts DMJ, Beets-Tan RGH (2021) Cross-sectional imaging for local staging of rectal cancer. In: Baatrup G (ed) *Multidisciplinary treatment of colorectal cancer: staging - treatment - pathology - palliation*. Springer International Publishing, Cham, pp 229-242 https://doi.org/10.1007/978-3-030-58846-5_24
6. Boldrini L, Charles-Davies D, Romano A et al (2025) Response prediction for neoadjuvant treatment in locally advanced rectal cancer patients-improvement in decision-making: A systematic review. *Eur J Surg Oncol* 51:109463 <https://doi.org/10.1016/j.ejso.2024.109463>
7. Pikūnienė I, Strakšytė V, Basevičius A et al (2023) Prognostic value of tumor volume, tumor volume reduction rate and magnetic resonance tumor regression grade in rectal cancer. *Medicina* 59:2194 <https://doi.org/10.3390/medicina59122194>
8. Lin YK, Zhu LL, Zhao J, Xiang ZL (2025) Radiotherapy volume delineation based on MRI and 18F-FDG-PET/MRI in locally recurrent rectal cancer. *Abdom Radiol (NY)* 50:4413-4422 <https://doi.org/10.1007/s00261-025-04859-2>
9. Joshi N, Bond S, Brady M (2010) The segmentation of colorectal MRI images. *Med Image Anal* 14:494-509 <https://doi.org/10.1016/j.media.2010.03.002>
10. Zhu W, Xu Y, Zhu H, Qiu B, Zhan M, Wang H (2024) Multi-parametric MRI radiomics for predicting response to neoadjuvant therapy in patients with locally advanced rectal cancer. *Jpn J Radiol* 42:1448-1457 <https://doi.org/10.1007/s11604-024-01630-3>
11. Jayaprakasam VS, Paroder V, Gibbs P et al (2022) MRI radiomics features of mesorectal fat can predict response to neoadjuvant chemoradiation therapy and tumor recurrence in patients with locally advanced rectal cancer. *Eur Radiol* 32:971-980 <https://doi.org/10.1007/s00330-021-08144-w>
12. Shaish H, Aukerman A, Vanguri R et al (2020) Radiomics of MRI for pretreatment prediction of pathologic complete response, tumor regression grade, and neoadjuvant rectal score in patients with locally advanced rectal cancer undergoing neoadjuvant chemoradiation: an international multicenter study. *Eur Radiol* 30:6263-6273 <https://doi.org/10.1007/s00330-020-06968-6>
13. Kaval G, Dagoglu Kartal MG, Azamat S et al (2024) Evaluating complete response prediction rates in locally advanced rectal cancer with different radiomics segmentation approaches. *Pathol Oncol Res* 30:1611744 <https://doi.org/10.3389/pore.2024.1611744>
14. Kwok HC, Charbel C, Danilova S et al (2022) Rectal MRI radiomics inter- and intra-reader reliability: should we worry about that? *Abdom Radiol (NY)* 47:2004-2013 <https://doi.org/10.1007/s00261-022-03503-7>
15. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp 234-241 https://doi.org/10.1007/978-3-319-24574-4_28
16. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18:203-211 <https://doi.org/10.1038/s41592-020-01008-z>

17. Ma J, Li F, Wang B (2024) U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 <https://doi.org/10.48550/arXiv.2401.04722>
18. Liu J, Yang H, Zhou HY et al (2024) Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp 615-625 https://doi.org/10.1007/978-3-031-72114-4_59
19. Soomro MH, Coppotelli M, Conforto S et al (2019) Automated segmentation of colorectal tumor in 3D MRI using 3D multiscale densely connected convolutional neural network. *J Healthc Eng* 2019:1075434 <https://doi.org/10.1155/2019/1075434>
20. Wang J, Lu J, Qin G et al (2018) Technical Note: A deep learning-based autosegmentation of rectal tumors in MR images. *Med Phys* 45:2560-2564 <https://doi.org/10.1002/mp.12918>
21. Filitto G, Coppola F, Curti N et al (2022) Automated prediction of the response to neoadjuvant chemoradiotherapy in patients affected by rectal cancer. *Cancers* 14:2231 <https://doi.org/10.3390/cancers14092231>
22. Knuth F, Groendahl AR, Winter RM et al (2022) Semi-automatic tumor segmentation of rectal cancer based on functional magnetic resonance imaging. *Phys Imaging Radiat Oncol* 22:77-84 <https://doi.org/10.1016/j.phro.2022.05.001>
23. Trebeschi S, van Griethuysen JJ, Lambregts DM et al (2017) Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci Rep* 7:5301 <https://doi.org/10.1038/s41598-017-05728-9>
24. Hamabe A, Ishii M, Kamoda R et al (2022) Artificial intelligence-based technology for semi-automated segmentation of rectal cancer using high-resolution MRI. *PLoS One* 17:e0269931 <https://doi.org/10.1371/journal.pone.0269931>
25. Zhang Z, Han J, Ji W et al (2024) Improved deep learning for automatic localisation and segmentation of rectal cancer on T2-weighted MRI. *J Med Radiat Sci* 71:509-518 <https://doi.org/10.1002/jmrs.794>
26. Yan W, Huang L, Xia L et al (2020) MRI Manufacturer Shift and Adaptation: Increasing the Generalizability of Deep Learning Segmentation for MR Images Acquired with Different Scanners. *Radiol Artif Intell* 2:e190195 <https://doi.org/10.1148/ryai.2020190195>
27. Knuth F, Askim Adde I, Huynh B et al (2022) MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts. *Acta Oncol* 61:255-263 <https://doi.org/10.1080/0284186X.2021.2013530>
28. Panic J, Defeudis A, Mazzetti S et al (2022) A fully automatic deep learning algorithm to segment rectal cancer on MR images: a multi-center study. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Glasgow, UK, pp 5066-5069 <https://doi.org/10.1109/embc48229.2022.9871326>
29. Panic J, Defeudis A, Vassallo L et al (2025) Rectal Cancer Segmentation: A Methodical Approach for Generalizable Deep Learning in a Multi-Center Setting. *Int J Imaging Syst Technol* 35:e70076 <https://doi.org/10.1002/ima.70076>
30. Fedorov A, Beichel R, Kalpathy-Cramer J et al (2012) 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* 30:1323-1341 <https://doi.org/10.1016/j.mri.2012.05.001>
31. Ibrahim RM, Said K (2017) Benefits & pitfalls of administration of rectal gel enema in MRI assessment of rectal cancer. *Egypt J Radiol Nucl Med* 48:847-852 <https://doi.org/10.1016/j.ejrn.2017.07.014>

32. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23:903-921 <https://doi.org/10.1109/TMI.2004.828354>
33. Pérez-García F, Sparks R, Ourselin S (2021) TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Programs Biomed 208:106236 <https://doi.org/10.1016/j.cmpb.2021.106236>
34. Ouyang C, Chen C, Li S et al (2023) Causality-Inspired Single-Source Domain Generalization for Medical Image Segmentation. IEEE Trans Med Imaging 42:1095-1106 <https://doi.org/10.1109/TMI.2022.3224067>
35. Arega TW, Legrand F, Bricq S, Meriaudeau F (2022) Using MRI-specific Data Augmentation to Enhance the Segmentation of Right Ventricle in Multi-disease, Multi-center and Multi-view Cardiac MRI. In: Puyol Antón E, Pop M, Martín-Isla C et al (eds) Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge. Springer International Publishing, Cham, pp 250-258 https://doi.org/10.1007/978-3-030-93722-5_27
36. Geißler K, Mensing D, Wenzel M, Hirsch JG, Heldmann S (2024) Towards TotalSegmentator for MRI data leveraging GIN data augmentation. In: Proc SPIE, vol 12926, p 1292604 <https://doi.org/10.1117/12.3006351>

Tables

Table 1. Summary of spatial resolution and acquisition characteristics across scanner vendors.

The table reports the mean values and ranges (in brackets) of the principal MRI acquisition parameters for each vendor. The dataset encompasses 19 scanner models in total, highlighting substantial variability in image dimensions, spatial resolution, and field of view across acquisition systems.

Vendor	No. of scanners	No. of subjects	Image width (px)	Image height (px)	Image depth (px)	Voxel size X (mm)	Voxel size Y (mm)	Voxel size Z (mm)	FOVx (mm)	FOVy (mm)	FOVz (mm)
Siemens	8	109	337.7 [212-640]	328.9 [212-640]	38.2 [22-60]	0.8 [0.3-1.3]	0.8 [0.3-1.3]	3.6 [3-6.6]	255 [166-400]	260 [166-400]	171.2 [78-322]
GE	5	62	468.7 [256-1024]	468.7 [256-1024]	32.8 [21-60]	0.5 [0.2-0.7]	0.5 [0.2-0.7]	3.5 [3-6]	277.4 [166-461]	234.1 [137-461]	163.3 [102-272]
Toshiba	1	31	648.3 [640-768]	650.1 [640-768]	31 [25-41]	0.4 [0.3-0.4]	0.4 [0.3-0.4]	3.9 [3.3-5]	252.7 [200-276]	265.7 [201-313]	189.1 [101-274]

Philips	4	23	507.1 [336-800]	507.1 [336-800]	36.7 [23-60]	0.5 [0.4-1.3]	0.5 [0.4-1.3]	4.4 [3-8]	264.4 [180-428]	279.8 [192-429]	201.9 [90-404]
Hitachi	1	1	512	512	24	0.5	0.5	6.5	281	281	159

Table 2. Scanner distribution and subject allocation across experimental setups.

Overview of scanner distribution and subject allocation across experimental setups. In the in-distribution configuration, all scanner vendors were represented in both training and testing subsets, ensuring domain familiarity between the two. In the out-of-distribution configuration, models were trained on images from Siemens, Toshiba and Philips scanners, and tested exclusively on scans acquired from GE and Hitachi systems.

MRI scanner	subjects (n)	In-distribution		Out-of-distribution	
		train (n)	test (n)	train (n)	test (n)
GE Optima MR450w	39	32	7	-	39
GE SIGNA Artist	4	3	1	-	4
GE SIGNA Creator	1	-	1	-	1
GE Signa Hdxt	12	10	2	-	12
GE SIGNA Voyager	6	5	1	-	6
Hitachi ECHELON	1	-	1	-	1
Philips Achieva	6	5	1	6	-
Philips Achieva dStream 1.5	10	8	2	10	-
Philips Achieva dStream 3	1	-	1	1	-
Philips Ingenia	6	5	1	6	-
Siemens Aera	12	10	2	12	-
Siemens Avanto	17	14	3	17	-
Siemens Avanto_fit	24	19	5	24	-
Siemens Biograph_mMR	9	7	2	9	-
Siemens Espree	11	9	2	11	-
Siemens Skyra	1	0	1	1	-
Siemens SymphonyTim	34	28	6	34	-
Siemens Verio	1	-	1	1	-
Toshiba Titan	31	26	5	31	-

total	226	181	45	163	63
-------	-----	-----	----	-----	----

Table 3. Summary of the augmentation transformations used during model training.

The table lists the data augmentation operations together with their sampling distributions and application probabilities.

Augmentation type	Transformation	Description	Application probability
base	rotation and scaling	Random affine transformations. Angles of rotation are drawn from $U(-180, 180)$; scaling factor is sampled from $U(0.7, 1.4)$.	0.2
	gaussian noise	Additive zero-centered noise perturbation. Variance of the noise is sampled from $U(0, 0.1)$.	0.15
	gaussian blur	Random isotropic smoothing. The width (in voxels) of the gaussian kernel is sampled from $U(0.5, 1.5)$.	0.2
	brightness	Voxel intensities are multiplied by $x \sim U(0.7, 1.3)$.	0.15
	contrast	Voxel intensities are multiplied by $x \sim U(0.65, 1.5)$ and then clipped to their original value range.	0.15
	simulation of low-resolution	Downsampling by a factor of $U(1, 2)$ using nearest neighbor interpolation and then sampling back up to the original size with cubic interpolation.	0.25
	gamma augmentation	Nonlinear intensity remapping with $\gamma \sim U(0.7, 1.5)$.	0.15
	mirroring	Random flipping along all spatial axes.	0.5
extended	bias field distortion	Reproduction of coil sensitivity and B1 inhomogeneity effects.	0.25
	random motion	Simulation of patient movement during acquisition.	0.25
	random ghosting	Introduction of periodic intensity duplication along the phase-encoding direction.	0.25
	Global Intensity Non-Linear augmentation	Nonlinear intensity alteration by using randomly-weighted shallow convolutional networks.	0.3

Table 4. In-distribution segmentation performance.

Mean Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) are reported for rectal cancer and mesorectum. Results compare the four ensemble configurations with and without MRI-specific data augmentation. Best values per metric are highlighted in bold.

	mean DSC $\pm$ std [range]	mean HD95 $\pm$ std [range] (mm)
--	----------------------------	----------------------------------

Ensembling configuration	Data augmentation	Rectal cancer	Mesorectum	Rectal cancer	Mesorectum
1	base	0.758 ± 0.119 [0.499-0.969]	0.774 ± 0.177 [0.023-0.967]	11.126 ± 11.299 [0.781-48.504]	9.059 ± 7.816 [0.781-35.642]
1	extended	0.766 ± 0.115 [0.540-0.966]	0.785 ± 0.159 [0.156-0.965]	11.536 ± 12.414 [0.781-46.319]	8.430 ± 6.702 [0.938-32.741]
2	base	0.756 ± 0.128 [0.500-0.967]	0.779 ± 0.160 [0.122-0.974]	12.924 ± 14.332 [0.781-60.610]	8.895 ± 6.877 [0.781-32.825]
2	extended	0.759 ± 0.125 [0.535-0.968]	0.794 ± 0.135 [0.393-0.976]	12.999 ± 14.547 [0.781-59.276]	8.416 ± 6.700 [0.781-32.750]
3	base	0.761 ± 0.120 [0.535-0.969]	0.779 ± 0.168 [0.075-0.971]	12.027 ± 13.159 [0.781-56.010]	8.561 ± 6.853 [0.781-32.877]
3	extended	0.767 ± 0.119 [0.552-0.971]	0.792 ± 0.138 [0.384-0.973]	12.078 ± 13.924 [0.781-59.263]	8.452 ± 6.609 [0.781-30.886]
4	base	0.751 ± 0.133 [0.479-0.964]	0.786 ± 0.138 [0.353-0.973]	14.073 ± 15.336 [0.781-61.622]	9.037 ± 7.058 [0.781-33.016]
4	extended	0.760 ± 0.125 [0.537-0.969]	0.794 ± 0.131 [0.394-0.977]	13.088 ± 15.047 [0.781-46.753]	8.516 ± 6.736 [0.394-32.832]

Table 5. Out-of-distribution segmentation performance.

Mean Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) are reported for rectal cancer and mesorectum. Results compare the four ensemble configurations with and without MRI-specific data augmentation. Best values per metric are highlighted in bold.

		mean DSC ± std [range]		mean HD95 ± std [range] (mm)	
Ensembling configuration	Data augmentation	Rectal cancer	Mesorectum	Rectal cancer	Mesorectum
1	base	0.790 ± 0.131 [0.219-0.969]	0.752 ± 0.194 [0.025-0.953]	6.957 ± 6.539 [0.586-27.566]	10.702 ± 10.119 [2.109-54.770]
1	extended	0.787 ± 0.136 [0.208-0.961]	0.766 ± 0.185 [0.118-0.965]	7.483 ± 8.422 [0.586-45.025]	9.944 ± 9.642 [0.829-51.883]
2	base	0.796 ± 0.125 [0.136-0.961]	0.770 ± 0.158 [0.258-0.968]	7.888 ± 7.904 [0.586-41.134]	9.414 ± 7.598 [0.586-34.534]
2	extended	0.792 ± 0.130 [0.136-0.964]	0.774 ± 0.162 [0.172-0.963]	7.797 ± 8.057 [0.586-46.586]	9.568 ± 8.671 [0.829-43.628]
3	base	0.797 ± 0.124	0.763 ± 0.178	6.959 ± 6.656	9.859 ± 8.604

		[0.153-0.968]	[0.116-0.964]	[0.586-31.208]	[1.172-39.880]
3	extended	0.792 ± 0.136 [0.153-0.965]	0.773 ± 0.171 [0.153-0.974]	7.563 ± 8.421 [0.442-44.928]	9.548 ± 8.980 [0.586-46.210]
4	base	0.798 ± 0.126 [0.111-0.960]	0.773 ± 0.157 [0.231-0.964]	7.583 ± 7.955 [0.586-41.024]	9.400 ± 7.904 [0.829-35.220]
4	extended	0.793 ± 0.129 [0.124-0.965]	0.776 ± 0.161 [0.175-0.967]	8.359 ± 9.043 [0.586-46.480]	9.501 ± 8.605 [0.586-43.657]

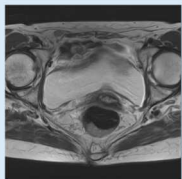
Figures legends

Figure 1 - The top row shows examples of four MRI scans, while the bottom row presents their corresponding manual contours (rectal tumor in blue and mesorectum in yellow). In cases A and B, the rectum was distended with ultrasound gel, providing clearer delineation of the rectal wall and tumor boundaries, whereas in C and D, no rectal filling was performed, resulting in partial lumen collapse and less distinct wall separation. The scanners used are: (A) Philips Achieva 1.5T, (B) Siemens SymphonyTim 1.5T, (C) GE Signa Artist 1.5T and (D) Toshiba Titan 3T.

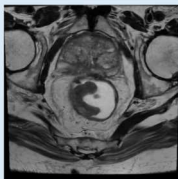
Figure 2 - Qualitative examples of RectoMap segmentations using the two-level STAPLE aggregation ensemble with extended MRI-specific augmentation. Panels A-B show in-distribution (ID) cases, while panels C-D illustrate out-of-distribution (OOD) cases. For each subject, the axial T2-weighted image (top), the expert manual ground truth (middle), and the model prediction (bottom) are displayed. Blue indicates the rectal tumor and yellow the mesorectum. The Dice similarity coefficients (DSC) for rectal tumor (RC) and mesorectum (MESO) are reported for each example.

IN-DISTRIBUTION

A

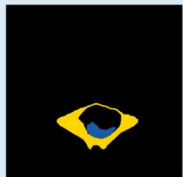


B



T2w image

ground truth

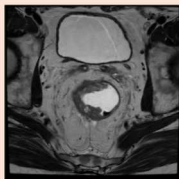


prediction

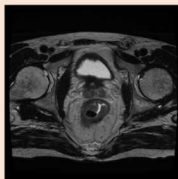


OUT-OF-DISTRIBUTION

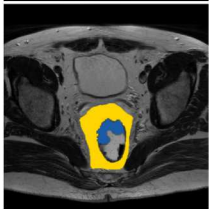
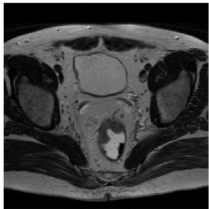
C



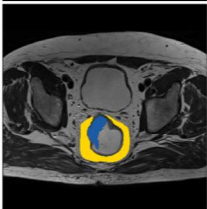
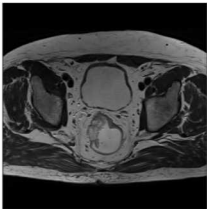
D



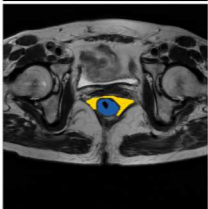
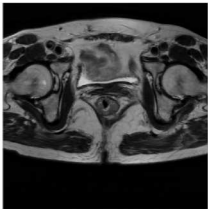
A



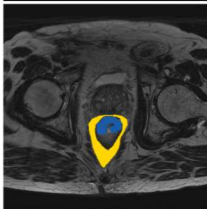
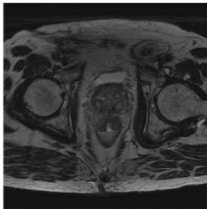
B



C



D



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryPERRA2026AbdomRadiol.docx](#)