Comparing Problem Formulations for Endometriosis Lesion Detection on T2-weighted Pelvic MRI

Eliot Leguy[1], Krystel Nyangoh-Timoh[1,2], Pierre Jannin[1], John S.H. Baxter[1], and Elodie Germani[1]

[1] Laboratoire Traitement du Signal et de l'Image (LTSI, INSERM UMR 1099), Université de Rennes, Rennes, France
[2] Centre Hospitalier Universitaire de Rennes (CHU Rennes), Rennes, France
Corresponding author: elodie.germani@univ-rennes.fr

Abstract. Deep infiltrative endometriosis (DIE), a subtype of endometriosis, is a chronic inflammatory disease where endometrial tissue (*i.e.*, lesions) grows outside the uterus and infiltrates pelvic organs such as the bladder or rectum. Accurate lesion mapping is crucial for diagnosis and surgical planning, notably for computing standardized severity scores such as the deep Pelvic Endometriosis Index (dPEI). Yet DIE lesions are small, low-contrast, and ill-delineated, making their detection on T2-weighted MRI challenging. On a retrospective, lesion-level-annotated cohort, we systematically compare four deep-learning formulations under a common instance-level protocol: dense segmentation (nnU-Net), bounding-box detection (YOLO), a latent token grid, and a patch-level classifier. We show that standard baseline formulations are insufficient, and the token-grid formulation yields only a partial, unstable spatial signal. The patch-level approach performs best, achieving a high local classification score in cross-validation but degrades significantly under full-slice sliding-window inference. An oracle analysis, in which windows are centered on lesions, shows that the local lesion signal is in fact highly separable. Taken together, these results indicate that the main bottleneck is not recognizing a lesion within a region, but generating anatomically relevant candidate regions across complete MRI slices. We argue that anatomically constrained candidate generation is the most promising direction towards a usable lesion mapping pipeline.

Keywords: Deep infiltrative endometriosis · Lesion detection · Pelvic MRI · Surgical planning.

1 Introduction

Endometriosis is a chronic, systemic, inflammatory, and hormone-dependent disease that affects around 10% of women of reproductive age worldwide [1]. Despite its high prevalence and substantial societal cost (estimated at €9.5 billion in 2025), the delay between symptom onset and diagnosis is 7 to 10 years, often

followed by years of therapeutic uncertainty [2]. Treatments aim to limit the impact of the typical symptoms (chronic pelvic pain, dysmenorrhea, and dyspareunia), but their perceived effectiveness remains mixed. On average, women change treatment 3.8 times over the course of their care pathway, and 62% of women undergo surgical excision of lesions [3]. These difficulties stem, among other factors, from the lack of reliable, objective techniques to diagnose the disease and to monitor its progression, currently relying on transvaginal ultrasound followed by pelvic MRI for diagnosis and on patient questionnaires for follow-up.

Endometriosis is traditionally divided into three subtypes: superficial endometriosis, the most common form, consisting of implants confined to the peritoneal surface; ovarian endometriosis, taking the form of cysts within the ovaries known as endometriomas; and deep infiltrative endometriosis (DIE), referring to lesions that infiltrate more than 5 mm beneath the peritoneal surface and may involve structures such as the uterosacral ligaments, the bladder or the rectum. DIE is strongly associated with severe pain, multi-organ involvement, and complex surgery, which makes lesion mapping particularly critical for surgical planning. To structure this assessment, standardized scores such as the deep Pelvic Endometriosis Index (dPEI) quantify compartmental involvement in the pelvis and help improve communication between radiologists and surgeons. Computing the dPEI, however, requires reliable localization of lesions within each anatomical compartment on T2-weighted pelvic MRI, which currently depends on time-consuming expert reading [5, 6].

Here, we aim to automate DIE lesion mapping on T2-weighted pelvic MRI. However, this task is challenging and does not fit the usual formulations in medical image analysis. First, DIE lesions are small, fibrous, and irregular, and often confounded with neighboring structures. Pixel-wise semantic segmentation, which is usually unstable when contours are ambiguous, could therefore fail on those structures. At the same time, the lesions are not well-contrasted, clearly delineated objects with spiculated non-standard shapes, so generic bounding-box detection might lack specificity relative to non-lesional pelvic structures. Previous work on automated MRI analysis for endometriosis has targeted patient-level diagnosis or lesion classification [13, 14], or segmentation of pelvic organs [15, 16], while radiomic methods mostly characterize lesions that are already delineated [17]. Even though lesion-level detection and mapping remain underexplored, studies such as [8] designed to segment pelvic structures, localize endometriosis plaques, and predict adhesions, suggest that lesion localization is more attainable than precise segmentation, particularly for DIE lesions, whose exact shape and volume are hard to define reproducibly.

Building on this observation, we frame DIE lesion mapping as a combined detection-and-localization problem rather than as fine-grained segmentation. Our contributions are threefold. First, we systematically compare four problem formulations for lesion mapping: dense semantic segmentation, bounding-box detection, a latent token grid that predicts lesions on a coarse spatial feature grid rather than at the pixel level, and a weakly supervised patch-level classifier. We evaluate these methods on a retrospective clinical cohort of 35 patients

with 83 manually annotated DIE lesions on T2-weighted MRI, under a common patient-level cross-validation and instance-level evaluation protocol. Second, we show that the patch-level formulation, combined with hard-negative mining, is the most promising approach, although its performance drops markedly when moving from local classification to full-slice sliding-window inference. Third, using an oracle analysis, we show that the local lesion signal is, in fact, highly separable, thereby identifying the main bottleneck in generating anatomically relevant candidate regions rather than in global detection.

2 Materials and Methods

2.1 Dataset

We use a retrospective, single-center clinical cohort of 35 patients who underwent surgery between 2020 and 2025 and had a T2-weighted pelvic MRI series available. The study was approved by the institutional review board of the CHU de Rennes, and all images were provided in anonymized DICOM format. Lesions were manually annotated by a radiology resident and exported as NRRD masks, each lesion carrying a unique identifier and an anatomical label. Across all series, 139 lesions were annotated, of which 84 corresponded to DIE. One DIE lesion was excluded because its annotation could not be consistently interpreted across the relevant slices, leaving 83 DIE lesions as the target for our task. The remaining 55 annotations (*e.g.*, adenomyosis, ovarian endometriomas, or hydrosalpinx) were not used as targets. The cohort comprises 3,193 T2 slices, of which 782 contain at least one DIE lesion. Acquisitions are markedly heterogeneous, with a mean slice thickness of 2.49 ± 1.51 mm and a mean in-plane resolution of 0.60 ± 0.18 mm, reflecting real clinical conditions and motivating the spatial standardization described below. Tab. 1 summarises the dataset.

2.2 Preprocessing

Each T2 volume and its annotation mask were reoriented to a common RAS orientation. Axial slices were resized to 320×320 pixels using linear interpolation followed by constant padding. Pixel intensities were normalized independently per volume by clipping to the $[0.5, 99.5]$ percentile range and rescaling to $[0, 1]$. Volumes were then converted to 2D axial slices, and each set of slices containing

Table 1. Dataset summary and preprocessed splits used for the main experiments. Positive slices contain at least one retained DIE lesion.

	Full cohort	Training	Test
Patients / T2 series	35	30	5
DIE lesions (retained)	83	74	9
Total slices	3,193	3,007	186
Positive slices	782	716	66
Positive-slice proportion	24.5%	23.8%	35.5%

at least one lesion voxel was identified and extended by 15 slices on each side, retaining anatomical context while discarding distant negative slices. Importantly, this lesion-centered extraction yields sets enriched for informative slices; thus, slice-level metrics should not be interpreted as an evaluation of full volumes.

2.3 Problem Formulations

For a preprocessed 2D slice we denote the input image $x \in \mathbb{R}^{1\times320\times320}$ and the binary DIE mask $m \in \{0,1\}^{320\times320}$. A slice is positive when it contains at least one lesional pixel, yielding a slice label $y = \mathbb{1}\big[\sum_{i,j} m_{i,j} > 0\big]$. The mask m further helps derive finer-grained spatial targets (presence maps, lesion centers, bounding boxes). We compare four formulations under this shared notation.

Dense semantic segmentation. As a biomedical segmentation reference, we train nnU-Net [9, 11] to predict per-pixel lesion probabilities, which are then thresholded into a binary mask whose connected components yield candidate instances. Both 2D and 3D full-resolution configurations are evaluated.

Bounding-box detection. As an object-detection reference, we train YOLOv8 [12], which predicts bounding boxes with confidence scores. Annotated lesion components are converted to minimal enclosing boxes for training, and DIE-free slices are included as negatives to expose the model to normal pelvic structures.

Latent token grid. This formulation predicts lesion location on a reduced spatial grid rather than per-pixel, assuming that DIE lesions occupy structured regions rather than isolated pixels. A variational autoencoder (VAE) f_θ, pretrained on public pelvic MRI [18], is used as a feature extractor. The encoder yields a latent feature grid $Z = f_\theta(x) \in \mathbb{R}^{C\times T\times T}$, on top of which a detection head h_ψ produces a local presence map $\hat{Y} \in [0,1]^{T\times T}$ and an auxiliary slice probability $\hat{y} \in [0,1]$. The spatial target is obtained by projecting m to the grid resolution, encoding a coherent presence region rather than exact contours. At inference, $\hat{Y}$ is thresholded, connected components are extracted, small regions are removed, and each remaining component becomes a candidate region.

Patch-level classifier. The fourth formulation learns a local classifier applied to windows on a regular grid. For each slice, we extract 128×128 context windows and define the labels only from the central 16×16 window. An example is positive if this central window contains at least n=5 lesional pixels. A negative example may contain lesional pixels in its periphery, providing hard negatives that force precise localization rather than neighborhood presence. Inputs are in a 2.5D representation (the central slice and its four neighbours $z-2, \ldots, z+2$). A residual 2D CNN with attention pooling produces a score $\hat{p}(c_y, c_x) \in [0,1]$ per central window. We use a two-phase training approach: a first phase on grid-sampled examples, followed by hard-negative mining (*i.e.*, false positives found by inference on the training set are reinjected in a second phase) to explicitly expose the model to anatomically ambiguous structures. During inference, the

classifier is applied in a sliding-window fashion over all valid centers. Scores are placed back into a spatial probability map, thresholded, and grouped into connected components to form candidate regions.

2.4 Evaluation

We use 5-fold cross-validation with patient-level splits to avoid data leakage, and hold out an independent test set of n=5 patients. For each formulation, the cross-validation models are applied separately to the test set, and metrics are averaged. Decision thresholds and post-processing parameters are selected on validation sets. The significant superiority of models is evaluated using unilateral Mann-Whitney rank-sum tests between metrics across folds.

We evaluate performance at two levels. Slice-level evaluation measures the ability to separate slices containing at least one DIE lesion from lesion-free slices, reported with the area under the ROC curve (AUC) and average precision (AP), together with precision, recall, and F_1. Instance-level evaluation directly measures lesion retrieval for mapping. Each annotated DIE lesion is considered a ground-truth instance, and model outputs are converted to candidate instances by thresholding and connected-component extraction. For map-producing methods (semantic segmentation and token grid), a predicted instance is matched to an annotated lesion if its Dice coefficient exceeds a threshold determined on the validation sets. For the patch-level model, a predicted component is a valid candidate if it covers at least one central window compatible with an annotated lesion. Unmatched lesions count as false negatives and spurious predictions as false positives. We report instance-level precision, recall, and F_1 jointly.

To disentangle lesion separability from spatial search, we run an oracle in which a classifier receives windows already centered on lesions and outputs whether this window contains a lesion. This analysis removes the candidate-search difficulty and provides an upper bound on local separability.

2.5 Implementation

Experiments are implemented in Python v3.10 (RRID: SCR_008394) with PyTorch v2.5.1 (RRID: SCR_018536) and MONAI v1.5.2 and run on a Linux Ubuntu 18.04.6 workstation with an Intel Core i9-9940X CPU, 128 GB RAM, and four NVIDIA Quadro RTX 5000 GPUs. For reference models, we use nnU-Netv2 [11] (RRID: SCR_028165) and Ultralytics YOLOv8 [12].

3 Results

3.1 Limited performance of reference baselines

Tab. 2 summarises the results of a representative model for each formulation. The semantic segmentation baselines demonstrate the worst performance: with the 2D nnU-Net, the mean Dice and IoU are effectively zero, and no lesions are

Table 2. Performance of the different formulations. Slice-level AUC and instance-level F_1 on the test set averaged ($\pm$ standard deviation) across models trained with 5-fold cross-validation (5CV). **Boldface** marks the best model, and underlined performance highlights the second best. * denotes significant superiority ($p < 0.05$) over the second best.

Formulation	Representative model	Slice AUC	Inst. F_1
Segmentation	nnU-Netv2 3D fullres	–	0.102 ± 0.034
Box detection	YOLOv8 2D	$\underline{0.683 \pm 0.060}$	0.140 ± 0.055
Token grid	Token det., 3 slices	0.463 ± 0.140	0.213 ± 0.082
Patch-level	Local inference	$\mathbf{0.943 \pm 0.032}^*$	$\mathbf{0.563 \pm 0.120}^*$
Patch-level	Sliding window	0.554 ± 0.054	$\underline{0.324 \pm 0.059}$

correctly recovered at the instance level. The 3D full-resolution configuration improves only marginally (mean Dice 0.065 ± 0.049, IoU 0.038 ± 0.030), detecting on average about one lesion per fold for an instance F_1 of 0.102 ± 0.034. The added volumetric context, therefore, helps, but pixel-accurate overlap remains a challenging objective for small, low-contrast, ambiguously bounded DIE lesions. Bounding-box detection seems better aligned with the task: YOLOv8 reaches a slice-level AUC of 0.683 ± 0.060 and AP of 0.513 ± 0.063, with an instance F_1 of 0.140 ± 0.055, confirming that approximate localization is more attainable than dense delineation. However, the model produces a large number of false positives (on average 21.2 per model fold). These baselines confirm the intrinsic difficulty of the task: dense segmentation is too complex for ambiguously bounded lesions, while standard box detection localizes approximately but lacks specificity.

3.2 Token-Grid formulation provides a partial but unstable signal

Predicting a lesion candidate region on a latent grid improves over dense segmentation but remains modest (Tab. 2, third row). Across variants, slice-level AUC stays close to chance (best 0.463 ± 0.140 for the 3-slice variant). Instance F_1 is improved compared to reference methods but does not exceed 0.213 ± 0.082, indicating that the global score derived from the lesion maps does not cleanly separate positive from negative slices. The token-grid thus learns a partial spatial signal, but its conversion into localized instances remains unstable.

3.3 Patch-level classification shows strong local lesion separability

The best performance is achieved with patch-level classification, which reaches an AUC of 0.943 ± 0.032 and an AP of 0.586 ± 0.152. At the instance level, the model shows a F_1 score of 0.563 ± 0.120. These results suggest that the model can discriminate between regions with and without lesions when the lesion is centered within a context window, thereby providing evidence of lesion signal.

The patch-level ablation over context- and center-window sizes (Tab. 3) shows that enlarging the spatial context substantially improves local classification. For a fixed 128×128 context window, the 16×16 center window maximizes AP and

Table 3. Patch-level ablation over context- and center-window sizes. Each cell reports mean AUC (top) and AP (bottom) over the five cross-validation folds. **Boldface** marks the best model. Performance differences between models were not significant.

Context \ Center	16	32	64	96
64	0.881 ± 0.022	0.898 ± 0.047	0.914 ± 0.044	–
	0.381 ± 0.097	0.407 ± 0.096	0.410 ± 0.099	–
96	0.920 ± 0.027	0.910 ± 0.026	0.910 ± 0.031	–
	0.521 ± 0.102	0.443 ± 0.065	0.461 ± 0.090	–
128	$\mathbf{0.943 \pm 0.032}$	0.911 ± 0.040	0.916 ± 0.032	0.876 ± 0.037
	$\mathbf{0.586 \pm 0.152}$	0.508 ± 0.134	0.449 ± 0.095	0.329 ± 0.053

precision, whereas a 32×32 center window yields the best recall: a smaller target enforces stricter localization and suppresses some false positives, while a larger one increases tolerance at the cost of specificity. This indicates that lesions do carry a signal that is exploitable when the candidate region is correctly defined.

However, when deployed in sliding-window inference, performance drops sharply. At the slice level, AUC falls to 0.554 ± 0.054 and AP to 0.411 ± 0.058, the slice-level F_1 is 0.538 ± 0.012, with high recall (0.852 ± 0.126) but low precision (0.401 ± 0.049). These results suggest that the model flags most positive slices but remains noisy. At the instance level, F_1 reaches 0.324 ± 0.059 (precision 0.296 ± 0.064, recall 0.395 ± 0.114). The model can propose a candidate region near an annotated lesion, but many lesions are still missed, and false positives persist. This gap between local and sliding-window instance mapping is crucial: it indicates that the bottleneck is not recognizing a lesion within a correctly centered region, but rather generating anatomically relevant candidates across the many ambiguous, imbalanced pelvic structures encountered when scanning a complete slice. The small number of patients and lesions and the heterogeneity of acquisitions plausibly contribute to this instability.

3.4 Oracle analysis localizes the bottleneck to candidate generation

To further investigate this, we train a simple oracle to classify 96×96 windows already centered on lesions. The model reaches an AP of 0.800 and an F_1 of 0.793, well above any sliding-window result. This confirms that the local lesion signal is highly separable once the region of interest is provided, and that the dominant difficulty lies in candidate generation rather than in local recognition. The oracle is an experimental upper bound, not a deployable detector, since it removes the spatial search that a clinical system must perform.

3.5 Qualitative results

A qualitative review of the patch-level predictions (Fig. 1) illustrates the three regimes behind these numbers: correct detections near annotated lesions, anatomical false positives on non-lesional pelvic structures, and missed small lesions. The recurring errors point to the priorities for future work: reducing anatomical false

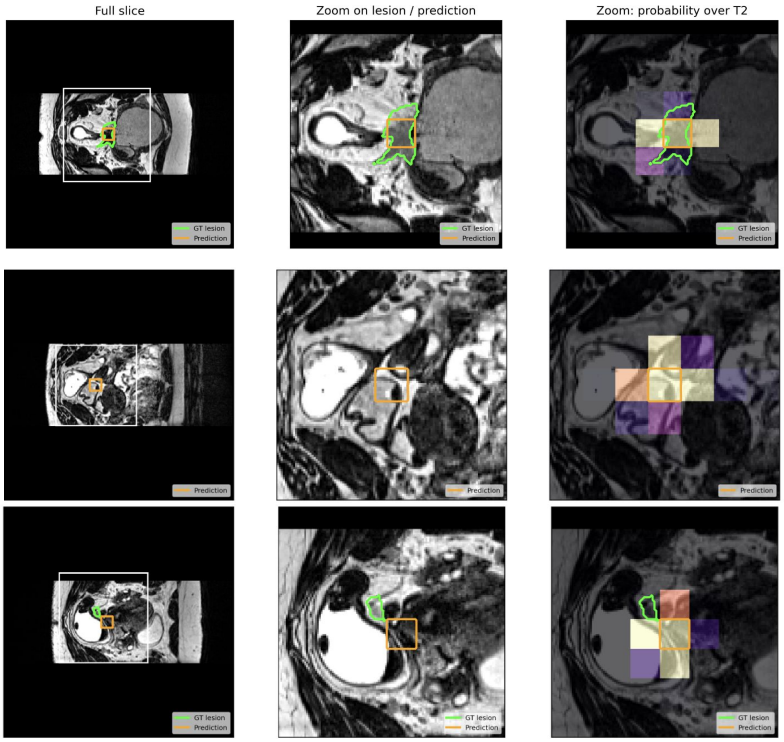

Fig. 1. Representative patch-level predictions on test slices, one regime per row. Columns: (left) full axial T2 slice with the analyzed region boxed; (center) zoom on the ground-truth lesion (green) and the predicted candidate (yellow); (right) the same zoom overlaid with the model's per-window probability map. Rows: (top) true positive, a predicted candidate overlapping an annotated lesion (central-window probability p = 0.807); (middle) false positive, a high-confidence prediction on a non-lesional pelvic structure (p = 0.736); (bottom) false negative, a small lesion assigned a sub-threshold score (p = 0.576). Probabilities are the score $\hat{p}(c_y, c_x)$ of the central 16×16 window.

positives, improving recall on small lesions, and constraining predictions with anatomical priors.

4 Conclusion

We framed automated DIE lesion mapping on T2-weighted pelvic MRI as a detection-and-localization problem and compared four formulations under a common patient-level, instance-level protocol. Dense segmentation and box detection proved insufficient, and the token grid gave only a partial, unstable signal. The patch-level classifier with hard-negative mining was clearly strongest locally but degraded under full-slice sliding-window inference. The oracle analysis shows the bottleneck lies in generating anatomically relevant candidates, not in local

recognition. These results stem from a single-center cohort of 35 patients and 83 lesions, with heterogeneous acquisitions, which limits their generalizability. Nonetheless, they point to anatomically constrained candidate generation, which leverages pelvic organ segmentation and anatomical priors, as the most promising path toward a clinically usable lesion-mapping pipeline.

References

1. World Health Organization: Endometriosis (2025), https://www.who.int/news-room/fact-sheets/detail/endometriosis, last accessed 2026/01/15
2. Chapron, C., Marcellin, L., Borghese, B., Santulli, P.: Rethinking mechanisms, diagnosis and management of endometriosis. Nature Reviews Endocrinology **15**, 666–682 (2019)
3. Candau, Y. & Estrade, D. Enquête EndoVie. *IPSOS*. (2020)
4. Bazot, M., et al.: Deep pelvic endometriosis: MR imaging for diagnosis and prediction of extension of disease. Radiology **232**, 379–389 (2004)
5. Thomassin-Naggara, I., et al.: Magnetic resonance imaging classification of deep pelvic endometriosis: description and impact on surgical management. Human Reproduction **35**, 1098–1108 (2020)
6. Reilhac, A., et al.: Validation of the deep pelvis endometriosis index (dPEI) to evaluate surgical outcomes of robotic-assisted surgery for endometriosis. Journal of Robotic Surgery **20**, 188 (2026)
7. van Waesberghe, J.H., Hazewinkel, M., Busard, M.: Endometriosis – MRI detection. The Radiology Assistant (2011)
8. Shiokawa, R., et al.: Development of an AI-based magnetic resonance imaging reading support program (AMP) for deep endometriosis diagnosis. Scientific Reports **16**, 790 (2025). https://doi.org/10.1038/s41598-025-30277-x
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
10. Isensee, F., et al.: nnU-Net Revisited: A call for rigorous validation in 3D medical image segmentation. In: MICCAI 2024, LNCS, vol. 15009, pp. 488–498. Springer (2024)
11. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Maier-Hein, K. H., nnU-Netv2, https://github.com/MIC-DKFZ/nnUNet/releases/tag/v2.0
12. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8 (2023), https://github.com/ultralytics/ultralytics
13. Moassefi, M., et al.: Advancing endometriosis detection in daily practice: a deep learning-enhanced multi-sequence MRI analytical model. Abdominal Radiology **51**(1), 238–249 (2026). https://doi.org/10.1007/s00261-025-04942-8
14. Zhang, L., Qing, X., Zou, J., Yang, H., Yu, L.: An ensemble machine learning model based on MRI features for diagnosing deep infiltrating endometriosis. Frontiers in Physiology **17** (2026). https://doi.org/10.3389/fphys.2026.1753847
15. Liang, X., et al.: A multi-modal pelvic MRI dataset for deep learning-based pelvic organ segmentation in endometriosis. Scientific Data **12**(1), 1292 (2025). https://doi.org/10.1038/s41597-025-05623-3
16. AlSaad, R., Albasha, S., Thomas, R.M.: Endo-MedSAM: a promptable vision foundation model adaptation for uterus segmentation on pelvic MRI in endometriosis. Frontiers in Reproductive Health **8** (2026). https://doi.org/10.3389/frph.2026.1790980

17. Peluso, S., et al.: Artificial intelligence-enhanced MRI radiomics for discriminating active and fibrotic lesions in deep endometriosis. Journal of Minimally Invasive Gynecology (2026). https://doi.org/10.1016/j.jmig.2025.12.037
18. Liang, X., Alpuing Radilla, L., Khalaj, K., Dawoodally, H., Mokashi, C., Guan, X., Roberts, K., Sheth, S., Tammisetti, V. & Giancardo, L. UTHealth - Endometriosis MRI Dataset (UT-EndoMRI). (Zenodo,2025,6)