A Staged Deep Learning Framework for Endometriosis-Aware Pelvic MRI Analysis Using UT-EndoMRI

Mursleen[1]*, Iftikhar Naseer[2],
[1]Faculty of Computer Science and Information Technology, Superior University, Lahore, 54600, Pakistan
[2]Superior College Mian Channu, 58000, Pakistan
***Correspondence:** mursasami@gmail.com

Citation | Mursleen., Naseer. I, "A Staged Deep Learning Framework for Endometriosis-Aware Pelvic MRI Analysis Using UT-EndoMRI", IJIST, Vol. 8 Issue.6 pp 2468-2482, September 2026
DOI:
Received | Aug 20, 2026 **Revised |** Aug 30, 2026 **Accepted |** Sep 05, 2026 **Published |** Sep 07, 2026

Endometriosis remains challenging to assess on pelvic MRI because lesions are often small, heterogeneous, and poorly distinguished from surrounding anatomy. This study developed a reproducible two-stage framework using the public UT-EndoMRI dataset to investigate whether lesion-aware segmentation can provide informative representations for subject-level endometriosis-aware prediction. In Stage 1, SegResNet, SwinUNETR, and SwinUNETR with lesion-aware refinement were compared for multi-class pelvic MRI segmentation. The refined SwinUNETR achieved the best test macro-Dice of 0.0786 and macro IoU of 0.0447, although ovary and lesion segmentation remained weak. The selected model was subsequently frozen and used to generate segmentation-informed features for Stage 2. A cohort of 121 subjects comprising 110 proxy/inferred positive and 11 proxy negative cases was evaluated using repeated five-fold stratified cross-validation over five repeats with L2-regularized logistic regression. The combined uncalibrated feature model achieved a balanced accuracy of 0.8215 ± 0.0798, MCC of 0.3953 ± 0.1172, and AUC of 0.8597 ± 0.0578. Subject-aggregated estimates were balanced accuracy 0.8364 (95% CI: 0.7955–0.8818), MCC 0.3968 (95% CI: 0.3407–0.4765), and AUC 0.8620 (95% CI: 0.7793–0.9306; permutation $p = 0.00498$). Isotonic calibration improved probability quality, reducing the Brier score to 0.0653 and ECE to 0.0380. These findings indicate that segmentation-informed representations can support internal subject-level prediction beyond coarse MRI summaries; however, the small negative class, weak lesion segmentation, proxy/inferred labels, possible target–predictor dependency, and absence of external validation preclude claims of clinical diagnostic validity.

Keywords: Deep Learning; Endometriosis; Pelvic Mri; Segmentation; Ut-Endomri

Introduction:

Problem Statement:

Endometriosis is a chronic condition with heterogeneous manifestations and a difficult diagnostic pathway. Contemporary clinical guidance recommends ultrasound or MRI as part of the diagnostic work-up while emphasizing that negative imaging does not exclude disease, particularly superficial peritoneal endometriosis [1]. For automated analysis, the challenge is also a measurement problem: lesions can be small, irregular, and poorly contrasted against adjacent pelvic structures, while normal pelvic anatomy varies substantially between subjects. Recent AI studies have reported promising endometriosis screening or classification performance using clinical symptoms, imaging, surgical video, and other data sources, but their endpoints, labels, and validation designs differ substantially[2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13]Consequently, strong performance in one modality cannot be assumed to transfer to multiparametric pelvic MRI or to clinically independent diagnosis.

Research Gap:

The public UT-EndoMRI resource provides multiparametric pelvic MRI together with pelvic-organ and lesion annotations, creating a rare opportunity for transparent, reproducible experimentation [14],[15]. Public data make preprocessing, partitioning, model selection, and downstream feature construction auditable rather than hidden within a private cohort. However, the resource does not remove label uncertainty: the Stage 2 labels available for this study were proxy or inferred rather than independently adjudicated clinical diagnoses. The present work therefore treats Stage 2 as an internal endometriosis-aware or lesion-presence prediction task and does not equate its results with diagnostic accuracy.

The unresolved gap is methodological rather than simply architectural. Recent endometriosis AI research includes symptom-based screening models [11][12]multimodal and image-based approaches [3][4][5][6][7][8][9][10]and broader reviews emphasizing phenotypic heterogeneity, reproducibility, and the need for rigorous validation [2][13]Meanwhile, modern medical-image segmentation has advanced from convolutional U-Net and nnU-Net pipelines toward transformer-based volumetric models such as SwinUNETR [16], [17] [18] [19] [20]What remains insufficiently studied is how an imperfect pelvic MRI segmentation representation behaves when it is frozen and reused for subject-level prediction, especially when label provenance, class imbalance, threshold selection, calibration, and uncertainty are made explicit. This study was designed to address that gap.

Literature Review:

Artificial intelligence research in endometriosis has expanded across clinical screening, ultrasound, magnetic resonance imaging (MRI), surgical imaging, ovarian-lesion differentiation, adenomyosis, and comorbidity-based prediction [3] [4] [5] [6] [7] [8] [9] [10] [11] [12]. developed machine-learning models for endometriosis screening and included prospective independent validation[11], whereas Goldstein and Cohen investigated symptom-based machine learning for earlier self-screening [12]. Imaging-oriented studies have similarly explored deep-endometriosis detection, ultrasound sliding signs, ovarian endometrioma discrimination, and computer-assisted surgical analysis [3] [4] [5] [6] [7] [8] [9] [10]Despite encouraging performance within individual studies, these approaches address heterogeneous clinical endpoints and use substantially different data sources, reference standards, cohort compositions, and validation strategies; therefore, their reported performance measures are not directly interchangeable. Recent reviews reinforce these limitations. [13] emphasized the heterogeneity of endometriosis phenotyping across surgery, omics, health records, and artificial intelligence. reviewed artificial-intelligence applications in endometriosis imaging and noted that the published literature remains dominated by ultrasound, with MRI-based investigations comparatively limited [21] A recent systematic review by Haber et al. further evaluated the methodological quality and diagnostic performance of AI approaches using

MRI, ultrasound, and patient-reported symptoms, highlighting the continuing need to distinguish promising technical performance from clinical readiness [22] Together, these findings support cautious interpretation of cross-study performance and reinforce the need for transparent validation and clearly defined target labels.

Pelvic MRI segmentation represents a particularly relevant component of this problem. The recently published UT-EndoMRI resource provides public multi-sequence pelvic MRI together with manual annotations of the uterus, ovaries, and detectable endometrioma regions and includes baseline automatic segmentation experiments [14] The dataset creates an important foundation for reproducible endometriosis-oriented MRI research while also illustrating the difficulty of segmenting small and anatomically variable pelvic structures. More broadly, biomedical image segmentation has progressed from the encoder–decoder architecture introduced by U-Net to self-configuring pipelines such as nnU-Net and transformer-based volumetric architectures such as SwinUNETR [16] [18] [17] [19] [20]. These advances provide increasingly capable tools for anatomical representation learning; however, improved segmentation methodology alone does not establish whether the resulting representations contain reliable subject-level information relevant to endometriosis. Accordingly, an unresolved methodological gap remains between pelvic MRI segmentation and subject-level endometriosis-aware prediction. Existing endometriosis AI studies primarily address direct screening, detection, classification, or segmentation tasks, whereas comparatively little evidence evaluates whether an independently frozen, imperfect segmentation representation can contribute useful subject-level information beyond global MRI summaries. This question becomes particularly important when label provenance, severe class imbalance, feature ablation, threshold selection, calibration, and statistical uncertainty are considered explicitly. The present study addresses this gap by linking a controlled Stage 1 pelvic MRI segmentation benchmark to frozen-feature Stage 2 classification and by evaluating the resulting representations using repeated stratified validation, calibration analysis, bootstrap uncertainty estimation, and permutation testing.

Research Objectives:

The objectives were directly derived from the identified gap. Objective 1 was to benchmark SegResNet, SwinUNETR, and SwinUNETR with lesion-aware refinement for multi-class pelvic MRI segmentation and quantify validation/test Dice and IoU. Objective 2 was to freeze the selected Stage 1 model and construct reproducible subject-level segmentation-informed features from its predictions. Objective 3 was to compare segmentation-only, MRI-only, combined, and lesion-volume-only feature sets under repeated stratified validation. Objective 4 was to quantify thresholded performance, ranking, calibration, negative-class sensitivity, bootstrap uncertainty, pairwise model differences, and permutation-based significance. The Results section follows these objectives in the same order so that each objective is traceable to a specific procedure and quantitative outcome.

Novelty Statement:

Unlike prior endometriosis AI studies that primarily predict disease from symptoms, ultrasound, surgical video, ovarian-cyst images, comorbidities, or other single-task inputs [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13], this study evaluates a controlled representation-to-classification pathway on public multiparametric pelvic MRI. The scientific advancement is not the introduction of a new backbone alone; it is the explicit measurement of how a frozen, imperfect segmentation representation contributes to subject-level prediction. The framework combines (i) a controlled three-model Stage 1 benchmark, (ii) lesion-aware refinement, (iii) frozen feature generation, (iv) four Stage 2 feature ablations, (v) provenance auditing for proxy/inferred labels, and (vi) repeated validation with fold-internal preprocessing, nested threshold selection, calibration, subject-level bootstrap confidence intervals, negative-class sensitivity analysis, and permutation testing. This design separates segmentation quality,

predictive discrimination, decision performance, and probability quality instead of reporting a single black-box accuracy score.

Materials and Methods:

Methodology Flow Diagram:

Figure 1 presents the complete analysis sequence: UT-EndoMRI preparation and subject-level splitting; orientation, resampling, normalization, four-channel MRI construction, and augmentation; Stage 1 training of SegResNet, SwinUNETR, and lesion-aware SwinUNETR; validation-based model selection; freezing of the selected Stage 1 model; generation of segmentation-informed and global MRI subject features; Stage 2 repeated stratified classification; fold-internal threshold selection and probability calibration; subject-aggregated evaluation; bootstrap uncertainty, negative-class sensitivity, and permutation analysis; and final interpretation. The diagram is retained because these requested elements are already represented by its staged blocks; the accompanying Materials and Methods text below now specifies the numerical configuration of each step.

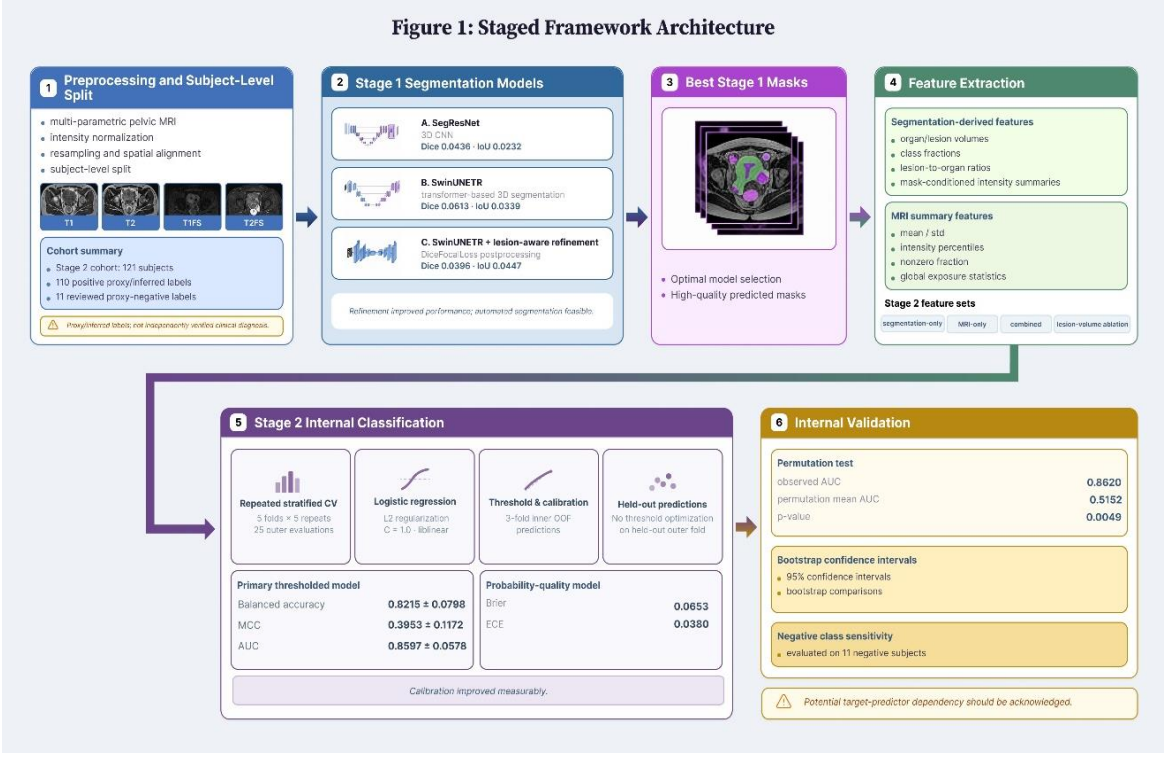

Figure 1. Overall architecture of the staged UT-EndoMRI framework

Dataset and Study Population:

The analysis used the public UT-EndoMRI dataset and the prepared UT-EndoMRI Prepared V1 derivative used in the Kaggle/Colab experiments [14][15]Each subject could contribute up to four MRI sequences—T1, T2, T1FS, and T2FS—which were represented as a consistent four-channel input. Missing sequences had been converted during dataset preparation to zero-filled placeholder volumes so that channel dimensionality remained fixed. Stage 1 used four labels: background (0), uterus (1), ovaries (2), and lesion/endometrioma region (3).

The prepared Stage 2 table contained 121 usable subjects: 110 proxy/inferred positive and 11 proxy negatives. The label-source audit identified 71 positives inferred from the D2_TCPW subset, 39 visible-lesion proxy positives, and 11 visible-lesion proxy negatives; no subject had an independently adjudicated clinical diagnosis label for Stage 2. Because visible-lesion information contributed to part of the target construction and predicted lesion masks also contributed to segmentation-informed predictors, target and predictor are not fully

independent. This dependency can inflate apparent association and is therefore treated as a central limitation. Stage 2 is interpreted only as internal endometriosis-aware/lesion-presence classification, not validated clinical diagnosis.

Table 1. Cohort and label-source audit for Stage 2

Label category	Subjects	Positive	Negative	Interpretation
D2_TCPW subset-inferred positive	71	71	0	Proxy/inferred
Visible-lesion proxy positive	39	39	0	Proxy
Visible-lesion proxy negative	11	0	11	Proxy
Independent clinical diagnosis label	0	0	0	Not available
Total usable cohort	121	110	11	Internal proxy-label cohort

Data Preprocessing:

Stage 1 preprocessing was reproduced from the experiment notebooks. Volumes were reoriented to RAS and resampled to 1.5 × 1.5 × 2.0 mm voxel spacing; MRI channels used bilinear interpolation and segmentation labels used nearest-neighbor interpolation. Each MRI sequence was normalized independently over non-zero voxels (channel-wise), after which T1, T2, T1FS, and T2FS were concatenated into a four-channel volume. Subject-level splitting with seed 42 produced 96 training subjects (87 positive, 9 negative), 12 validation subjects (11 positive, 1 negative), and 13 test subjects (12 positive, 1 negative), preventing subject leakage across partitions. Training augmentation included random flips on two spatial axes (p=0.5 each) and random affine augmentation (p=0.15; rotation ranges 0.08, 0.08, 0.04 radians; scale ranges 0.05, 0.05, 0.03; translation up to 4, 4, 2 voxels). Lesion-aware crops were sampled from a binary lesion mask. Evaluation used deterministic preprocessing without augmentation.

Proposed Method:

The study was organized as two connected stages. Stage 1 asked whether architecture strength and lesion-aware optimization improved multi-class pelvic MRI segmentation. After validation-based selection, the strongest Stage 1 model was frozen; no Stage 2 optimization was allowed to update its weights. Stage 2 then summarized each subject using features derived from predicted masks and MRI intensities. The authoritative journal-validation feature definitions contained 48 segmentation-informed features (predicted class volumes/fractions, lesion-to-organ ratios, and mask-conditioned modality statistics), 20 global MRI summary features (mean, standard deviation, 10th/90th percentiles, and non-zero fraction for each modality), 64 combined features after removing duplicate global non-zero terms, and a 2-feature lesion-volume-only ablation (predicted lesion voxels and lesion fraction). This design directly tested whether the frozen segmentation representation contributed information beyond coarse global MRI summaries.

Table 2. Stage 2 feature-set definitions

Feature set	No. of features	Purpose
Combined	64	Segmentation-derived variables plus MRI summary statistics
Segmentation-only	48	Predicted volumes, ratios, fractions, and mask-conditioned intensity summaries
MRI-only	20	Global modality means, standard deviation, percentiles, and nonzero fraction
Lesion-volume-only	2	Minimal ablation using predicted lesion voxels and lesion fraction

Model Architecture:

SegResNet was the lightweight 3D convolutional baseline (four input channels, four output classes, 16 initial filters, dropout 0.15). SwinUNETR was the transformer-based

comparison (four input channels, four output classes, feature size 24, 3D spatial configuration, gradient checkpointing enabled). The refinement experiment retained the SwinUNETR backbone and changed only the optimization strategy by increasing lesion-positive crop sampling and replacing standard Dice loss with DiceFocalLoss (Dice weight 0.7; focal weight 0.3; background excluded). The selected Stage 1 model was frozen as a representation generator rather than treated as a clinically reliable lesion delineator, because lesion and ovary Dice remained low.

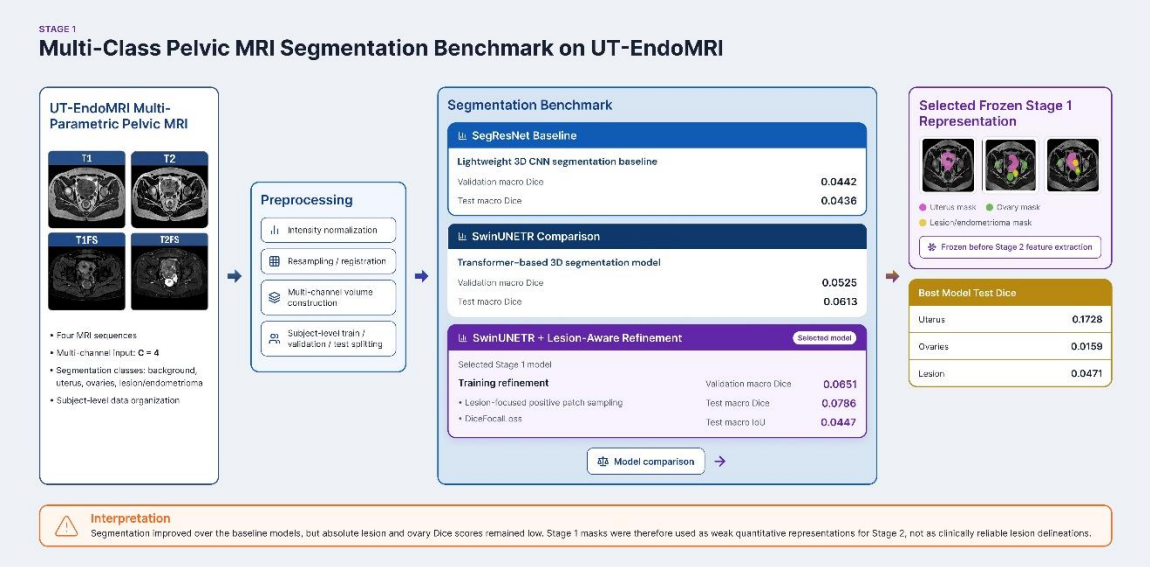

Figure 2. Stage 1 multi-class pelvic MRI segmentation benchmark.

For Stage 2, logistic regression was selected because the cohort was small and highly imbalanced and because the primary question concerned the information carried by the frozen Stage 1 representation. Logistic regression estimated the positive-class probability by applying a sigmoid transformation to a linear score.

Training Configuration:

All Stage 1 runs used seed 42, AdamW optimization (learning rate 2×10^{-4}; weight decay 1×10^{-5}), batch size 1, 15 epochs, mixed-precision training, and sliding-window inference with batch size 1 and 0.25 overlap. SegResNet used 64 × 64 × 48 voxel training/inference patches; SwinUNETR and its refinement used 64 × 64 × 64 patches to satisfy the transformer hierarchy. Standard SegResNet and SwinUNETR training used Dice loss with background excluded and lesion-positive:negative crop sampling of 3:1. The refinement run used a 4:1 lesion-positive: negative crop ratio and DiceFocalLoss while keeping the SwinUNETR architecture unchanged. The checkpoint with the highest validation macro-Dice was retained for final validation/test reporting; the held-out test partition was not used for model selection.

Stage 2 used repeated stratified five-fold cross-validation with five repeats (25 outer evaluations; seed 42). Within every outer-training split, missing values were imputed with a constant, features were standardized, and minority-class random oversampling was applied only to training observations. An L2-regularized logistic-regression model (C=1.0, liblinear solver, maximum 2000 iterations) was then fitted. A three-fold inner out-of-fold procedure on the outer-training data generated probabilities for operating-threshold selection, preventing optimization on the held-out outer fold. The threshold maximized balanced accuracy, with MCC used to resolve ties. Sigmoid and isotonic calibration were fitted using inner out-of-fold training probabilities rather than held-out labels. This nesting kept preprocessing, oversampling, threshold selection, and calibration inside the training portion of each outer evaluation.

Evaluation and Statistical Analysis:

Objective-specific evaluation used Dice and IoU for Stage 1 and balanced accuracy, MCC, AUC, Brier score, and expected calibration error (ECE) for Stage 2. Balanced accuracy and MCC were emphasized because only 11 negative subjects were available. Threshold-dependent metrics and threshold-independent ranking were reported separately, and calibration was evaluated with Brier score/ECE rather than inferred from AUC. Analytical definitions are provided in the Results section. Reporting choices were additionally reviewed against contemporary medical-AI reporting and validation guidance [23][24], [25][26] Subject-aggregated summaries averaged each subject's repeated held-out probabilities and corresponding fold thresholds to provide one post hoc internal summary per subject; these aggregates are not an independent test set. A one-sided Monte Carlo permutation test evaluated whether the subject-aggregated AUC exceeded chance after label randomization (200 valid permutations). The finite permutation counts limits p-value resolution to $1/(200+1)=0.00498$ for the smallest non-zero result.

Stratified subject-level bootstrap confidence intervals were calculated with 3000 resamples using the mean operating threshold of each subject-aggregated model. Paired stratified bootstrap comparisons evaluated the combined model against MRI-only, lesion-volume-only, and segmentation-only settings. Negative-class sensitivity was also reported explicitly as counts among the 11 negative subjects. These analyses quantify internal uncertainty but cannot compensate for the small negative class, proxy/inferred targets, or absence of external validation.

Results:

Evaluation Measures and Analytical Formulation:

The quantitative results reported below use the following definitions. For foreground class c, TP, FP, and FN denote true-positive, false-positive, and false-negative voxels. Macro segmentation scores average the class-specific values across the evaluated foreground classes.

$$\mathrm{Dice}(c) = \frac{2\mathrm{TP}_c}{2\mathrm{TP}_c+\mathrm{FP}_c+\mathrm{FN}_c} \quad (1)$$

$$\mathrm{IoU}(c) = \frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c+\mathrm{FN}_c} \quad (2)$$

$$\mathrm{MacroMetric} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{Metric}(c) \quad (3)$$

For Stage 2, p_i denotes the estimated positive-class probability for subject i. The operating threshold τ* was selected by maximizing balanced accuracy on inner out-of-fold predictions, with MCC used to resolve ties.

$$p_i = \mathrm{P}(y_i = 1 \mid x_i) = \frac{1}{1+\exp\left(-(\beta_0+\beta^T x_i)\right)} \quad (4)$$

$$\tau^* = \arg\max_{\tau}\frac{1}{2}[\mathrm{TPR}(\tau)+\mathrm{TNR}(\tau)] \quad (5)$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right) \quad (6)$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{[(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})]}} \quad (7)$$

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(p_i-y_i)^2 \quad (8)$$

$$\mathrm{ECE} = \sum_{b=1}^{B}\frac{|S_b|}{n}|\mathrm{acc}(S_b)-\mathrm{conf}(S_b)| \quad (9)$$

$$p_{\mathrm{perm}} = \frac{1+\sum_{r=1}^{R}\mathrm{I}(\mathrm{AUC}_{\mathrm{perm},r}\geq\mathrm{AUC}_{\mathrm{obs}})}{R+1} \quad (10)$$

Segmentation Benchmark:

The first study aim was addressed through a controlled comparison of SegResNet, SwinUNETR, and SwinUNETR with lesion-aware refinement. A consistent improvement in macro segmentation performance was observed from the convolutional baseline to the

transformer-based model and again after lesion-aware refinement. SegResNet achieved a test macro-Dice of 0.043597 and macro IoU of 0.023237, whereas standard SwinUNETR increased these values to 0.061281 and 0.033917, respectively. The refined SwinUNETR produced the strongest overall performance, with validation macro-Dice of 0.065078 and IoU of 0.036797 and test macro-Dice of 0.078596 and IoU of 0.044732. These results indicate that both model architecture and lesion-aware optimization influenced segmentation performance. However, the absolute metric values remained modest, demonstrating that multi-class pelvic MRI segmentation in this dataset remained challenging despite the observed improvements.

Table 3. Final Stage 1 segmentation comparison

Model	Val macro-Dice	Val macro IoU	Test macro-Dice	Test macro IoU
SegResNet	0.044193	0.024065	0.043597	0.023237
SwinUNETR	0.052514	0.028807	0.061281	0.033917
SwinUNETR+ refinement	0.065078	0.036797	0.078596	0.044732

Per-class analysis further clarified the limitations of the selected Stage 1 model. Uterus segmentation was comparatively the most stable, achieving a test Dice of 0.172804 and IoU of 0.101048. In contrast, ovary segmentation remained particularly difficult, with a test Dice of 0.015928, while lesion segmentation achieved a test Dice of 0.047056. The refined model should therefore be interpreted as the strongest representation generator among the evaluated Stage 1 models, rather than as a clinically reliable anatomical or lesion delineation system.

Table 4. Per-class Dice and IoU for the selected Stage 1 model

Class	Validation Dice	Validation IoU	Test Dice	Test IoU
Uterus	0.137858	0.079981	0.172804	0.101048
Ovaries	0.007832	0.003957	0.015928	0.008096
Lesion	0.049544	0.026451	0.047056	0.025052

This limitation was retained explicitly when interpreting Stage 2 rather than being obscured by downstream classification performance.

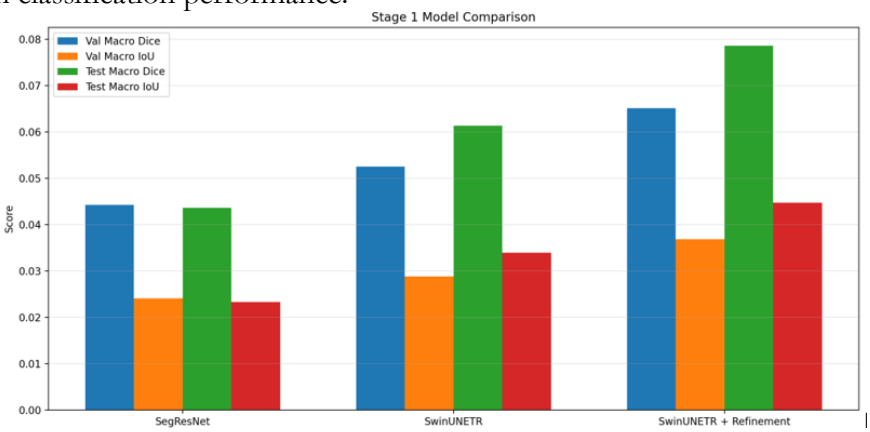

Figure 3. Stage 1 model comparison across validation and test macro-Dice and IoU.

Frozen Feature Construction and Stage 2 Comparison:

Following Stage 1 model selection, the refined SwinUNETR was frozen and used solely to generate subject-level segmentation-informed representations. Its weights were not updated during Stage 2. The resulting segmentation-derived variables were evaluated independently and in combination with global MRI summary features to determine whether the frozen Stage 1 representation contributed information beyond coarse intensity statistics. Repeated stratified cross-validation showed that the combined uncalibrated model provided the strongest threshold-dependent classification performance, achieving balanced accuracy of 0.8215 ± 0.0798 and MCC of 0.3953 ± 0.1172, with an AUC of 0.8597 ± 0.0578. The segmentation-only representation produced an AUC of 0.8639 ± 0.0676, which was marginally

higher than that of the combined model, although its balanced accuracy and MCC were lower. This finding indicates that the frozen segmentation representation contained substantial information for the constructed Stage 2 target, while also showing that the combined representation was not superior on every metric. MRI-only features performed less strongly, with balanced accuracy of 0.7352 ± 0.1363 and AUC of 0.7924 ± 0.1137. The lesion-volume-only model showed balanced accuracy of 0.5000 and MCC of 0.0000, indicating that predicted lesion burden alone was insufficient to account for the Stage 2 classification results.

Table 5. Key Stage 2 repeated cross-validation results

Role	Feature set	Model	Bal. acc.	MCC	AUC	Brier	ECE
Primary thresholded	Combined	Uncalibrated	0.8215 ± 0.0798	0.3953 ± 0.1172	0.8597 ± 0.0578	0.1254 ± 0.0425	0.1663 ± 0.0560
Probability quality	Combined	Isotonic calibrated	0.8079 ± 0.0951	0.3804 ± 0.1341	0.8512 ± 0.0834	0.0686 ± 0.0159	0.0594 ± 0.0384
Segmentation ablation	Segmentation-only	Uncalibrated	0.7418 ± 0.1229	0.2982 ± 0.1612	0.8639 ± 0.0676	0.1264 ± 0.0411	0.1634 ± 0.0498
MRI-only ablation	MRI-only	Uncalibrated	0.7352 ± 0.1363	0.2752 ± 0.1710	0.7924 ± 0.1137	0.1910 ± 0.0533	0.2351 ± 0.0673
Lesion-burden ablation	Lesion-volume-only	Sigmoid calibrated	0.5000 ± 0.0000	0.0000 ± 0.0000	0.6030 ± 0.1449	0.0830 ± 0.0116	0.0307 ± 0.0100

Calibration and Subject-Aggregated Internal Evaluation:

Calibration analysis was conducted separately from discrimination and threshold-dependent classification because these quantities address different aspects of model performance. When repeated held-out predictions were aggregated to one mean prediction per subject, the combined uncalibrated model retained the strongest thresholded behavior, with balanced accuracy of 0.8364, MCC of 0.3968, and AUC of 0.8620.

In contrast, isotonic calibration substantially improved probability quality. For the combined model, the subject-aggregated Brier score decreased from 0.1177 for the uncalibrated model to 0.0653, while ECE decreased from 0.1233 to 0.0380. Its subject-aggregated AUC remained similar at 0.8612. These findings demonstrate a clear distinction between classification performance at a selected operating threshold and probability calibration; therefore, no single Stage 2 configuration is considered universally optimal.

For thresholded internal classification, the combined uncalibrated model is the preferred configuration, whereas the isotonic-calibrated combined model is preferable when probability reliability is the primary consideration.

Table 6. Subject-aggregated Stage 2 validation results

Role	Feature set	Model	Bal. acc.	MCC	AUC	Brier	ECE
Primary thresholded	Combined	Uncalibrated	0.8364	0.3968	0.8620	0.1177	0.1233
Probability quality	Combined	Isotonic calibrated	0.8091	0.3733	0.8612	0.0653	0.0380
Segmentation ablation	Segmentation-only	Uncalibrated	0.8136	0.3810	0.8628	0.1194	0.1461

Role	Feature set	Model	Bal. acc.	MCC	AUC	Brier	ECE
MRI-only ablation	MRI-only	Uncalibrated	0.7773	0.3248	0.7603	0.1857	0.2379
Lesion-burden ablation	Lesion-volume-only	Sigmoid calibrated	0.5000	0.0000	0.5347	0.0832	0.0190

The subject-aggregated confusion matrices are shown in Figure 4. They are reported separately from Table 6 to keep the table readable and to avoid mixing compact performance summaries with cell-level count information.

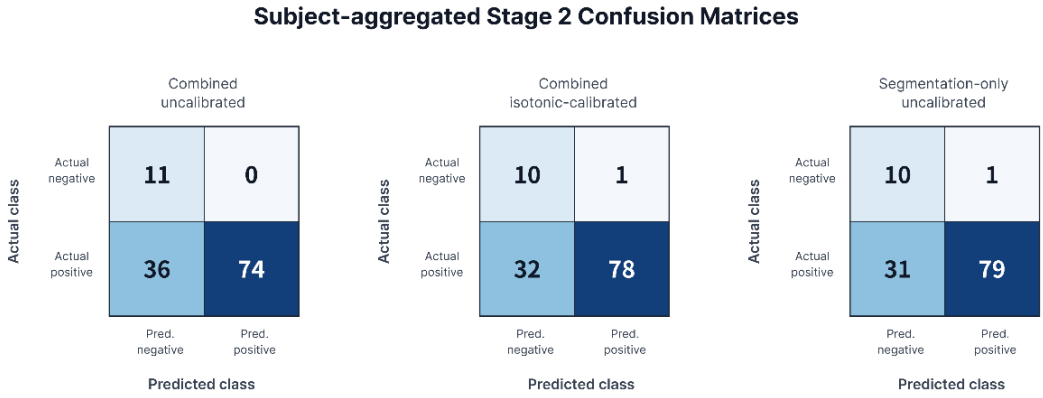

Figure 4. Subject-aggregated confusion matrices for selected Stage 2 models

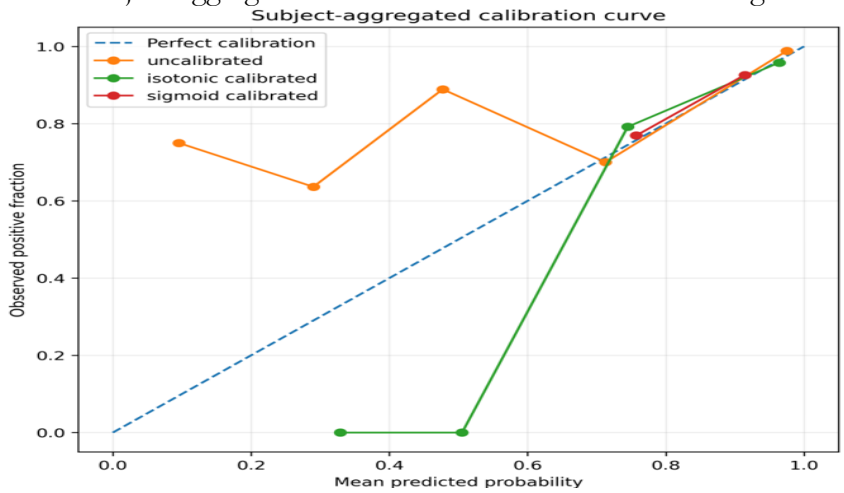

Figure 5. Subject-aggregated calibration curves for the combined Stage 2 models.

Statistical Robustness and Uncertainty Analyses:

The statistical analyses completed Objective 4. The observed subject-aggregated AUC of 0.8620 exceeded the permutation mean AUC of 0.5152; with 200 valid permutations, the one-sided p-value was 1/201=0.00498. This supports above-chance internal ranking under the implemented pipeline but does not establish external validity, clinical utility, or causal independence between predictors and targets.

For the combined uncalibrated model, stratified subject-level 95% bootstrap CIs were 0.7955–0.8818 for balanced accuracy, 0.3407–0.4765 for MCC, and 0.7793–0.9306 for AUC. Compared with MRI-only features, combined features increased AUC by 0.1024 (95% CI 0.0207–0.1884); compared with lesion-volume-only features, they increased balanced accuracy by 0.3371, MCC by 0.4001, and AUC by 0.3267, with intervals excluding zero. Pairwise intervals did not clearly separate combined and segmentation-only settings. Among only 11 negative subjects, the combined uncalibrated model classified 11/11 correctly, versus 10/11

for combined isotonic, 9/11 for segmentation-only, 10/11 for MRI-only, and 0/11 for lesion-volume-only. These negative-class counts are descriptive and statistically fragile because a one-subject change corresponds to approximately 9.1 percentage points of specificity.

Discussion:

Interpretation:

Stage 1 produced a consistent model ordering, with lesion-aware SwinUNETR performing best, but the absolute Dice values remained low, particularly for ovaries and lesions. The selected model should therefore be interpreted as the strongest representation generator in this experiment rather than as a clinically reliable delineation system.

Stage 2 showed that segmentation-informed features carried useful signal for the constructed internal target: the segmentation-only AUC was competitive and the combined feature set gave the strongest thresholded performance. The lesion-volume-only ablation remained at chance for balanced accuracy and MCC, indicating that lesion burden alone did not explain the result. However, because visible-lesion information contributed to proxy-label construction as well as segmentation-informed predictors, target-feature dependency remains a potential source of optimistic association.

Calibration produced a clear trade-off. Isotonic calibration improved Brier score and ECE, whereas the uncalibrated combined model retained the strongest thresholded classification. Thus, the preferred output depends on whether the analysis prioritizes discrete research classification or probability quality.

Comparison with Previous Research:

Direct numerical superiority claims over previous endometriosis AI studies are inappropriate because recent work spans symptom-based screening, ultrasound, surgical video, ovarian-cyst discrimination, adenomyosis, comorbidity data, and other endpoints [3] [4] [5] [6] [7] [8] [9][10] [11] [12] [13]. For example, recent symptom-based studies achieved strong internal or cohort-specific discrimination [11] [12] [13]but their input data and target definitions differ fundamentally from the present proxy-labeled multiparametric MRI task. The contribution here is therefore methodological: a public MRI segmentation benchmark is linked to a frozen representation-to-classification experiment with explicit ablations and uncertainty analysis. This extends the medical-segmentation literature [27] [18] [17] [19] [20] while following recent recommendations that AI evaluation distinguish discrimination, calibration, robustness, and intended clinical use [23] [24],[25] [26]

Clinical / Practical Significance:

The practical value of the framework is as an auditable research pipeline for pelvic MRI rather than a deployable diagnostic system. First, Stage 1 quantifies which pelvic targets the current models can and cannot localize; the particularly low ovary and lesion Dice values identify where improved annotation, higher-resolution acquisition, class-aware loss design, or stronger 3D models are most needed. Second, the Stage 2 ablations show that spatially conditioned representations can retain subject-level information even when segmentation is imperfect, which supports further research on representation-assisted MRI analysis. Third, the calibration and uncertainty analyses demonstrate why a high AUC alone is insufficient for a small, imbalanced cohort. In a future clinically labeled setting, a similar staged design could support radiologists by providing structured quantitative features or research decision support, but only after independently adjudicated labels, external/site-level validation, prospective evaluation, and direct comparison with expert readers. These requirements are consistent with current endometriosis guidance and medical-AI evaluation recommendations [1][25]

Strengths:

Methodological strengths include the use of a public multiparametric MRI resource, subject-level partitioning, a controlled Stage 1 benchmark, a frozen representation for Stage 2, explicit feature-set ablations, repeated stratified cross-validation, fold-internal preprocessing,

calibration analysis, subject-level bootstrap uncertainty, negative-class sensitivity analysis, and a permutation test. The label-source audit and explicit distinction between internal methodological performance and clinical diagnostic validity further reduce the risk of overstating the findings.

Limitations:

The principal statistical limitation is the small, highly imbalanced cohort (110 positive and 11 negative subjects). Repeated stratified cross-validation, balanced metrics, and subject-level bootstrap intervals reduce dependence on one split but cannot create information that is absent from the negative class. Specificity, calibration, and threshold stability are therefore particularly sensitive to individual negative subjects; for example, one classification change among 11 negatives changes specificity by approximately 9.1 percentage points.

Stage 2 labels were not independently adjudicated diagnoses. Seventy-one positives were inferred from the D2_TCPW subset, while 39 positives and 11 negatives were constructed using visible-lesion proxy rules. Because predicted lesion masks and lesion-conditioned intensity feature also enter the Stage 2 predictor set, target–predictor dependency is possible. The observed associations may therefore be optimistic for a true independent diagnostic task. The lesion-volume-only ablation, which remained at chance for balanced accuracy and MCC, shows that a single lesion-burden variable did not by itself explain the full result, but it does not eliminate the dependency concern.

Segmentation was weak for ovaries and lesions. Although the downstream classifier may still benefit from the resulting representations, the predicted masks should not be used as clinical lesion delineations.

There is no external validation or prospective clinical comparison. Future work should use larger and more balanced pelvic MRI cohorts with independently adjudicated endometriosis labels, prespecified external-site validation, prospective evaluation, and comparison against expert radiologist interpretation. Cross-site scanner/protocol robustness, subgroup performance, confidence-interval width, calibration transportability, and clinically meaningful operating thresholds should be assessed before any diagnostic claim. Reproducibility should include the Stage 1 split/training configuration, model/checkpoint identifiers, Stage 2-fold assignments, processed feature definitions, label-source audit, and held-out predictions.

Conclusion:

This paper presented a staged deep learning framework for endometriosis-aware pelvic MRI analysis using UT-EndoMRI. Objective-wise evaluation showed that lesion-aware SwinUNETR was the strongest Stage 1 representation generator, that its frozen segmentation-informed features carried subject-level signal, and that combining segmentation-informed and MRI features provided the strongest thresholded Stage 2 behavior under the implemented internal validation. These findings are methodological rather than diagnostic because Stage 1 lesion/ovary segmentation remained weak and Stage 2 labels were proxy/inferred.

The combined uncalibrated Stage 2 model achieved repeated-CV balanced accuracy 0.8215 ± 0.0798, MCC 0.3953 ± 0.1172, and AUC 0.8597 ± 0.0578; subject-aggregated AUC was 0.8620 (95% CI 0.7793–0.9306; permutation $p=0.00498$). Isotonic calibration improved Brier score and ECE, demonstrating a trade-off between the strongest internal thresholded decision rule and the best probability quality. The framework should therefore be viewed as a transparent experimental pipeline for future validation, not as a clinically deployable endometriosis diagnostic model.

Future studies should prioritize independently adjudicated labels, substantially larger negative cohorts, external multi-site validation, prospective evaluation, protocol/scanner robustness testing, and reader-level comparison with radiologists. Only after these steps should the clinical utility of segmentation-informed pelvic MRI prediction be assessed.

Data Availability:

The public UT-EndoMRI dataset is available through the cited dataset records [14][15]Code, environment details, Stage 1 split/training configuration, Stage 2 fold assignments, processed features, label audit, and held-out predictions are available at: https://share.google/gq3V0zQITNeNucvvC

Acknowledgement:

The authors acknowledge the creators of the publicly available UT-EndoMRI resource. The authors also confirm that this manuscript has not been published previously and is not under consideration by another journal.

Author's Contribution:

All authors contributed significantly to the work, reviewed the manuscript, agree with its content, and approve its submission.

Conflict of Interest:

The authors declare no conflict of interest.

Project Details / Funding:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References:

[1] C. M. Becker *et al.*, "ESHRE guideline: endometriosis," *Hum. Reprod. open*, vol. 2022, no. 2, 2022, doi: 10.1093/HROPEN/HOAC009.

[2] B. Sivajohan, M. Elgendi, C. Menon, C. Allaire, P. Yong, and M. A. Bedaiwy, "Clinical use of artificial intelligence in endometriosis: a scoping review," *NPJ Digit. Med.*, vol. 5, no. 1, p. 109, Dec. 2022, doi: 10.1038/S41746-022-00638-1.

[3] S. Guerriero *et al.*, "Artificial intelligence (AI) in the detection of rectosigmoid deep endometriosis," *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 261, pp. 29–33, Jun. 2021, doi: 10.1016/j.ejogrb.2021.04.012.

[4] G. Maicas *et al.*, "Deep learning to diagnose pouch of Douglas obliteration with ultrasound sliding sign," *Reprod. Fertil.*, vol. 2, no. 4, p. 236, Oct. 2021, doi: 10.1530/RAF-21-0031.

[5] K. Miao, Q. Lv, L. Zhang, N. Zhao, and X. Dong, "Discriminative diagnosis of ovarian endometriosis cysts and benign mucinous cystadenomas based on the ConvNeXt algorithm," *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 298, pp. 135–139, Jul. 2024, doi: 10.1016/j.ejogrb.2024.05.010.

[6] H. Wang *et al.*, "Human-AI collaborative multi-modal multi-rater learning for endometriosis diagnosis," *Phys. Med. Biol.*, vol. 70, no. 1, Jan. 2024, doi: 10.1088/1361-6560/AD997E.

[7] A. Bondarenko *et al.*, "Object Detection in Laparoscopic Surgery: A Comparative Study of Deep Learning Models on a Custom Endometriosis Dataset," *Diagnostics 2025, Vol. 15, Page 1254*, vol. 15, no. 10, p. 1254, May 2025, doi: 10.3390/DIAGNOSTICS15101254.

[8] A. Hernández *et al.*, "Deep Learning to Measure the Intensity of Indocyanine Green in Endometriosis Surgeries with Intestinal Resection," *J. Pers. Med.*, vol. 12, no. 6, Jun. 2022, doi: 10.3390/JPM12060982.

[9] D. Raimondo *et al.*, "Application of Deep Learning Model in the Sonographic Diagnosis of Uterine Adenomyosis," *Int. J. Environ. Res. Public Health*, vol. 20, no. 3, Feb. 2023, doi: 10.3390/IJERPH20031724.

[10] U. Tore *et al.*, "Diagnosis of Endometriosis Based on Comorbidities: A Machine

Learning Approach," *Biomedicines*, vol. 11, no. 11, Nov. 2023, doi: 10.3390/BIOMEDICINES11113015.

[11] S. Bendifallah *et al.*, "Machine learning algorithms as new screening approach for patients with endometriosis," *Sci. Reports 2022 121*, vol. 12, no. 1, pp. 639-, Jan. 2022, doi: 10.1038/s41598-021-04637-2.

[12] A. Goldstein and S. Cohen, "Self-report symptom-based endometriosis prediction using machine learning," *Sci. Reports 2023 131*, vol. 13, no. 1, pp. 5499-, Apr. 2023, doi: 10.1038/s41598-023-32761-8.

[13] C. R. Nezhat *et al.*, "Real world perspectives on endometriosis disease phenotyping through surgery, omics, health data, and artificial intelligence," *npj Women's Heal. 2025 31*, vol. 3, no. 1, pp. 8-, Feb. 2025, doi: 10.1038/s44294-024-00052-w.

[14] X. Liang *et al.*, "A Multi-Modal Pelvic MRI Dataset for Deep Learning-Based Pelvic Organ Segmentation in Endometriosis," *Sci. Data 2025 121*, vol. 12, no. 1, pp. 1292-, Jul. 2025, doi: 10.1038/s41597-025-05623-3.

[15] X. Liang *et al.*, "UTHealth - Endometriosis MRI Dataset (UT-EndoMRI)", doi: 10.5281/ZENODO.13749613.

[16] Olaf Ronneberger, Philipp Fischer, Thomas Brox., "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv*, vol. 1, 2015, doi: https://doi.org/10.48550/arXiv.1505.04597.

[17] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12962 LNCS, pp. 272–284, Jan. 2022, doi: 10.1007/978-3-031-08999-2_22.

[18] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods 2020 182*, vol. 18, no. 2, pp. 203–211, Dec. 2020, doi: 10.1038/s41592-020-01008-z.

[19] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[20] M. J. Cardoso *et al.*, "MONAI: An open-source framework for deep learning in healthcare," Nov. 2022, Accessed: Sep. 06, 2026. [Online]. Available: https://arxiv.org/pdf/2211.02701

[21] S. Mittal, A. Tong, S. Young, and P. Jha, "Artificial intelligence applications in endometriosis imaging," *Abdom. Radiol. 2025 5010*, vol. 50, no. 10, pp. 4901–4913, Apr. 2025, doi: 10.1007/S00261-025-04897-W.

[22] M. Haber, M. Montebello, G. P. Gauci, F. Zarb, and K. B. Grima, "Bridging the Gap Between Artificial Intelligence and Clinical Readiness in Endometriosis Diagnosis: A Systematic Review," *J. Minim. Invasive Gynecol.*, vol. 33, no. 9, pp. 1242–1250, Sep. 2026, doi: 10.1016/j.jmig.2026.04.016.

[23] A. S. Tejani *et al.*, "Checklist for Artificial Intelligence in Medical Imaging (CLAIM):2024 Update," *https://doi.org/10.1148/ryai.240300*, vol. 6, no. 4, May 2024, doi: 10.1148/RYAI.240300.

[24] G. S. Collins *et al.*, "TRIPOD+AI statement: Updated guidance for reporting clinical

prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, Apr. 2024, doi: 10.1136/BMJ-2023-078378.

[25] B. Vasey *et al.*, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nat. Med. 2022 285*, vol. 28, no. 5, pp. 924–933, May 2022, doi: 10.1038/s41591-022-01772-9.

[26] L. Maier-Hein *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nat. Methods 2024 212*, vol. 21, no. 2, pp. 195–212, Feb. 2024, doi: 10.1038/s41592-023-02151-z.

[27] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597*, 2015, [Online]. Available: https://arxiv.org/abs/1505.04597

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.