Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction of Lymph Node Metastasis Risk in Endometrial Cancer Using Multi-Sequence Magnetic Resonance Imaging Combined with Deep Transfer Learning

Tao Zheng
The First Hospital of Qinhuangdao
Fei Hu
The First Hospital of Qinhuangdao
Dan Du
The First Hospital of Qinhuangdao
Xin Liang
The First Hospital of Qinhuangdao
Juan Du
The First Hospital of Qinhuangdao
Tian Ye
The First Hospital of Qinhuangdao
Jiangyang Pan
The Fourth Hospital of Hebei Medical University
Linsha Yang
The First Hospital of Qinhuangdao
Qinglei Shi
School of Medicine, Chinese University of Hong Kong (Shenzhen) & Shenzhen Research Institute of Big Data
Defeng Liu
defengliu20230163.com

The First Hospital of Qinhuangdao

Research Article

Keywords: endometrial cancer, lymph node metastasis, magnetic resonance imaging, radiomics, deep transfer learning

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10972182/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose: To develop and externally validate a nomogram integrating clinical predictors, multisequence magnetic resonance imaging (MRI) radiomics, and deep learning (DL) features for preoperative lymph node metastasis (LNM) prediction in endometrial cancer (EC), and assess the incremental value of DL.

Methods: This two-center retrospective study included 520 patients: 268 for training, 117 for held-out internal validation, and 135 for external validation. Radiomics features were extracted from four MRI sequences, and DL features from four ImageNet-pretrained 50-layer residual network (ResNet-50) branches fine-tuned in training. The Clinical+Radiomics model combined four clinical predictors with a radiomics signature; the nomogram additionally incorporated the DL signature. Final models were fitted using all training patients and evaluated unchanged in both validation cohorts. Assessments included DeLong tests, calibration, decision curve analysis, and prespecified sensitivity analysis in magnetic resonance (MR) report-based LNM-negative patients.

Results: In internal and external validation, respectively, the nomogram achieved areas under the receiver operating characteristic curve (AUCs) of 0.866 and 0.886, with 95% confidence intervals (CIs) of 0.786–0.946 and 0.817–0.955. Corresponding Clinical+Radiomics AUCs were 0.733 (95% CI: 0.591–0.875) and 0.715 (95% CI: 0.594–0.836), favoring the nomogram (P=0.024 and P<0.001, respectively). Brier scores were 0.110 and 0.083. Decision curves favored the nomogram over comparators at probability thresholds of 0.02–0.39 and 0.01–0.71, respectively. Nomogram AUCs in MR report-based LNM-negative subgroups were 0.843 and 0.858, respectively.

Conclusion: Adding DL improved discrimination beyond Clinical+Radiomics. The nomogram may complement preoperative LNM risk stratification but should not replace pathological nodal staging. Prospective multicenter validation is required before clinical implementation.

1. Introduction

Endometrial cancer (EC) is a common gynecologic malignancy with rising incidence and mortality. Lymph node metastasis (LNM) is a major adverse prognostic factor that influences surgical strategy, adjuvant therapy, and recurrence risk[1–4]. Although systematic lymphadenectomy provides definitive nodal staging, routine use in presumed early-stage EC remains controversial because randomized trials have not shown clear survival benefit and the procedure increases operative time and complications such as lymphocele and lower-limb lymphedema[5–10]. Therefore, reliable preoperative prediction of LNM is needed to identify patients who may benefit from nodal assessment while sparing low-risk patients unnecessary intervention.

Magnetic resonance imaging (MRI) is widely used preoperatively for EC, as it assesses myometrial invasion, cervical stromal spread, and extrauterine extension. Yet conventional nodal evaluation on MRI, computed tomography (CT), or positron emission tomography/computed tomography (PET/CT) still mainly relies on size, shape, and metabolic activity, which misses micrometastatic disease and metastases in morphologically normal nodes. As a result, pathology remains the standard for definitive LNM diagnosis[11]. Multiparametric MRI provides richer information, but subtle spatial heterogeneity across sequences is not always captured by visual reading.

MRI radiomics can extract high-dimensional quantitative descriptors of lesion intensity, shape, texture, and heterogeneity[12–15]. Prior studies suggest radiomics-based models can improve preoperative LNM assessment, especially when combined with clinical data. However, handcrafted features depend on image acquisition, reconstruction, segmentation, and feature-selection workflows, and the high feature-to-event ratio may compromise stability and increase overfitting risk in limited positive-event cohorts[16, 17]. Conventional statistical and shallow machine-learning methods may also fail to capture complex nonlinear imaging patterns.

Deep learning (DL), particularly convolutional neural networks (CNNs), can learn hierarchical representations directly from images and may capture information beyond handcrafted features[18–20]. Transfer learning further allows pretrained models to be adapted to relatively small medical datasets, improving convergence and reducing computational demand[21, 22]. Yet DL models still face concerns about interpretability, overfitting, and cross-institution generalizability. Combining clinical variables with both handcrafted radiomics and DL transfer-learning features may provide complementary information and improve robustness.

Accordingly, this study aimed to develop and validate a nomogram integrating clinical variables, multisequence MRI radiomics, and deep transfer-learning features for preoperative LNM prediction in EC. Model performance was tested in internal hold-out and independent external cohorts, including discrimination, calibration, and potential clinical utility.

2. Materials and methods

2.1 Patient inclusion and cohort allocation

This retrospective two-center study enrolled consecutive patients with EC treated at Center A (XXX Hospital of XXX) and Center B (XXX Hospital of XXX) from January 2013 to June 2024. The Ethics Committee of Center A approved the study, and informed consent was waived because all data were anonymized.

Eligible patients had histologically confirmed EC after total hysterectomy with bilateral salpingo-oophorectomy, available pathological lymph-node assessment by sentinel lymph-node biopsy (SLNB) and/or lymphadenectomy, and preoperative MRI including axial fat-suppressed T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) maps, and delayed-phase three-dimensional volumetric interpolated breath-hold examination (3D-VIBE). Exclusion criteria were severe MRI artifacts (n = 11), lesions too small or invisible on MRI (n = 84), MRI-to-surgery interval > 4 weeks (n = 41), and preoperative radiotherapy or chemotherapy (n = 60). Complete data were available for all preoperative clinical predictors in the included patients; therefore, no imputation was required.

Of 716 screened patients, 520 were included. Center A contributed 385 patients, randomly split at the patient level into training and internal validation cohorts (7:3) before model development; all images from one patient remained in the same cohort. The internal validation cohort was set aside before model development and used exclusively for performance evaluation. Center B provided 135 patients for external validation. Patient flow is shown in Supplementary Fig. S1.

2.2 Pathological assessment

Nodal assessment was performed by SLNB, systematic lymphadenectomy, or both. Of the 520 patients, 130 (25.0%) underwent SLNB alone, 286 (55.0%) underwent systematic lymphadenectomy alone, and 104 (20.0%) underwent both procedures. Sentinel-node tracers included indocyanine green, carbon nanoparticles, and methylene blue. Side-specific systematic lymphadenectomy was performed when sentinel-node mapping failed. Among patients undergoing SLNB alone, para-aortic nodes were pathologically assessed only when para-aortic sentinel nodes were successfully mapped and excised; otherwise, no para-aortic nodal specimen was available for pathological evaluation.

Systematic pelvic and para-aortic lymphadenectomy was additionally performed according to institutional surgical practice and preoperative risk assessment, especially in patients with high-grade or non-endometrioid tumors, deep myometrial invasion (DMI), or suspected nodal disease, consistent with contemporary recommendations for surgical lymph-node staging[23].

Nodes removed by systematic lymphadenectomy were evaluated by routine hematoxylin and eosin (H&E) staining. SLNB nodes negative on initial H&E underwent ultrastaging with serial sectioning and cytokeratin immunohistochemistry[24]. The same criteria for nodal tumor-deposit classification and LNM status were applied at both centers.

Nodal deposits were categorized as macrometastases (> 2 mm), micrometastases (> 0.2–2 mm), or isolated tumor cells (ITCs; ≤0.2 mm). For the binary pathological reference standard, LNM-positive status required at least one node with macro- or micrometastasis. ITCs were recorded as pN0(i+) and not counted as LNM-positive. Patients without macro- or micrometastases in the examined lymph nodes were classified as LNM-negative. Unsampled nodal regions were not considered pathologically confirmed negative. Postoperative pathology slides and reports were reviewed retrospectively by a pathologist with > 5 years of experience, blinded to clinical and imaging data. Histologic subtype and grade, depth of myometrial invasion, and nodal status were extracted. All patients were restaged using the 2023 International Federation of Gynecology and Obstetrics (FIGO) staging system[25].

2.3 Extraction of variables from preoperative MRI reports

Magnetic resonance (MR) report-based DMI, maximum tumor diameter (MTD), and LNM status were extracted directly from the original preoperative clinical MR reports. None of these variables was reassigned through retrospective image review.

DMI was defined as invasion of at least 50% of the myometrial thickness[26]. MTD was recorded in centimeters as documented in the original report, without retrospective remeasurement.

These MR report-based variables were used in baseline comparisons and clinical modeling. MR report-based LNM status additionally defined the prespecified sensitivity-analysis subgroups.

2.4 MRI acquisition

MRI was performed using a 1.5-T scanner at Center A and a 3.0-T scanner at Center B. The protocol included axial fat-suppressed T2WI, DWI with ADC maps, and multiphase contrast-enhanced 3D-VIBE. Delayed-phase images were acquired approximately 180 s after contrast administration. Detailed parameters are provided in Supplementary Table S1.

2.5 Tumor segmentation

T2WI, high-b-value DWI, ADC, and delayed-phase 3D-VIBE images were exported from the picture archiving and communication system (PACS) and converted from Digital Imaging and Communications in Medicine (DICOM) to Neuroimaging Informatics Technology Initiative (NIfTI). Using ITK-SNAP v3.4.0, two radiologists (one from each center) manually delineated two-dimensional (2D) tumor regions of interest (ROIs) on all tumor-containing axial T2WI slices, including necrotic and hemorrhagic areas; T2WI was the reference sequence. Segmentation was performed independently at each center with blinding to clinical and pathological data.

For reproducibility, 30 patients from the training cohort were randomly selected, and both radiologists repeated T2WI tumor delineation after 1 month. The workflow is shown in Fig. 1.

2.6 Image preprocessing

All four MRI sequences were corrected with N4 bias-field correction implemented in Advanced Normalization Tools (ANTs). T2WI was used as the reference for segmentation and spatial alignment. DWI, ADC, and 3D-VIBE were rigidly registered to T2WI, and the T2WI-derived tumor mask was propagated to each aligned sequence.

For radiomics, registered images and propagated masks were resampled to sequence-specific resolutions: T2WI and 3D-VIBE at 0.9 × 0.9 × 4 mm^3, and DWI and ADC at 1.2 × 1.2 × 4 mm^3, using linear interpolation for images and nearest-neighbor interpolation for binary masks.

For DL, aligned DWI, ADC, and delayed-phase 3D-VIBE were kept in T2WI space. Tumor-centered bounding boxes from the T2WI reference mask were applied to all four sequences to generate ROI crops. Each crop was resized to 224 × 224 pixels and z-score normalized per sequence and per tumor-containing slice using statistics from that individual crop (mean and standard deviation), not across patients or cohorts. Tumor masks were used only to localize ROI crops and were not used for intensity normalization or as network input channels. This image-level normalization was therefore independent of downstream feature or signature standardization.

2.7 Radiomics feature extraction

2D radiomics features were extracted from each tumor-containing slice with PyRadiomics v2.2.0 following the Image Biomarker Standardization Initiative (IBSI) recommendations. Included feature families were first-order, shape, gray-level co-occurrence, gray-level size-zone, gray-level run-length, neighboring gray-tone difference, and gray-level dependence descriptors.

Preprocessing transformations included wavelet, Laplacian-of-Gaussian (LoG; σ = 1.0 and 5.0 mm), square, square-root, logarithm, exponential, gradient, and 2D local binary pattern filters. A total of 1,589 features were extracted per sequence. Slice-level features were aggregated to the patient level by tumor-area-weighted averaging before feature selection and model fitting.

Volumetric filters (including wavelet) were applied to the three-dimensional (3D) image volume before slice-wise radiomics computation; texture was then calculated on axial slices. Three-letter wavelet codes (e.g., HLL, LHL, HHH) therefore denote 3D decomposition directions, not 3D texture analysis.

Feature reproducibility was assessed with intra- and interobserver intraclass correlation coefficients (ICCs); only features with ICC > 0.75 in both analyses were retained. Detailed extraction settings, patient-level feature aggregation, and reproducibility assessment are provided in Supplementary Methods S1.

2.8 DL and feature extraction

A four-branch multisequence deep transfer-learning model used parallel 50-layer residual network (ResNet-50) branches. Branches processed ROI crops from T2WI, DWI, ADC, and 3D-VIBE, sharing the same architecture but independent parameters. All branches used ImageNet-pretrained weights and were jointly fine-tuned for LNM classification in the training cohort. Each branch produced 2,048 features via global average pooling; vectors were concatenated per aligned slice. Hyperparameter tuning, augmentation, and early stopping were confined to training.

After fine-tuning, the fusion head was removed, and the fixed model output slice features. Borderline Synthetic Minority Oversampling Technique (Borderline-SMOTE), feature standardization, Pearson correlation coefficient (PCC) filtering, least absolute shrinkage and selection operator (LASSO) selection, and logistic regression generated slice-level LNM probabilities. Tumor-area-weighted soft voting yielded a patient-level probability, then logit-transformed and z-score normalized as the DL signature. The classification threshold was fixed in training and kept for internal and external validation. Further network and training details are provided in Supplementary Methods S2.

2.9 Feature selection and model development

Model training used five-fold patient-level out-of-fold (OOF) validation stratified by LNM, with all data from each patient kept within one fold. In each training fold, clinical variables were selected by backward Akaike information criterion (AIC) selection, and radiomics and DL signature pipelines were rebuilt with Borderline-SMOTE, standardization, PCC/LASSO, and logistic regression[27]. Detailed oversampling, nested feature-selection, and signature-standardization procedures are provided in Supplementary Methods S3. The Clinical+Radiomics model included retained clinical predictors and the radiomics signature, while the nomogram additionally included the DL signature. OOF-derived optimal thresholds (Youden index) were fixed, then final models were refit on all 268 training patients and applied unchanged to both validation cohorts. No additional downstream tuning or early stopping was performed beyond inner LASSO cross-validation. Clinical predictor selection, combined-model development, and OOF evaluation are detailed in Supplementary Methods S4.

ResNet-50 fine-tuning was not repeated in OOF, so OOF reflected downstream modeling on fixed representations. Held-out internal and external cohorts provided primary performance evaluation, and gradient-weighted class activation mapping (Grad-CAM) was used only for exploratory interpretation[28].

2.10 Sensitivity analysis

A prespecified sensitivity analysis was conducted among patients with negative MR report-based LNM status in the held-out internal and external validation cohorts. MR report-based LNM status was determined from the original preoperative clinical MR reports. The locked model developed in the training cohort was applied without refitting, recalibration, feature reselection, or threshold optimization. Performance was evaluated separately in the two validation cohorts using the prespecified training-derived threshold.

2.11 Statistical analysis

Statistical analyses were performed using MedCalc version 20.0.14 and Python 3.9.25. Continuous variables were compared using independent-samples *t* tests, one-way analysis of variance, or Mann–Whitney U tests, as appropriate, and categorical variables using chi-square or Fisher exact tests.

Model discrimination was assessed using receiver operating characteristic (ROC) and precision–recall (P-R) curves, with their respective areas under the curve (AUCs). ROC AUCs were compared using the DeLong test. P-R AUC point estimates were calculated by trapezoidal integration of precision with respect to recall, based on patient-level predicted probabilities. The 95% confidence intervals (CIs) for P-R AUC were obtained using 1000 nonparametric patient-level bootstrap resamples (random seed, 978), with the P-R AUC recalculated in each bootstrap sample. For the training cohort, performance estimates were based on pooled patient-level OOF predictions; the internal and external validation cohorts were evaluated using predictions generated by the locked final models.

Calibration was assessed using calibration curves, Brier scores, calibration intercepts and slopes, and the Hosmer–Lemeshow goodness-of-fit test; detailed calibration procedures are provided in Supplementary Methods S5. Clinical utility was evaluated using decision curve analysis (DCA) over threshold probabilities from 0.00 to 0.80. All tests were two-sided, and $P < 0.05$ was considered statistically significant.

3. Results

3.1 Clinical and pathological characteristics

The training, internal validation, and external validation cohorts included 268, 117, and 135 patients, respectively (Table 1). Body mass index (BMI), reproductive history, and MR report-based DMI differed significantly across cohorts (P=0.032, P=0.024, and P<0.001, respectively). Age, menopausal status, cancer antigen 125 (CA125), human epididymis protein 4 (HE4), MR report-based MTD, MR report-based LNM status, histological type and grade, lymphovascular space invasion, FIGO stage, and pathological LNM status did not differ significantly (all P>0.05).

3.2 Feature selection and development of the clinical model

AIC-based backward selection (AIC=198.809) retained HE4, MR report-based DMI, MR report-based MTD, and MR report-based LNM status in the full training cohort. The clinical model yielded AUCs of 0.765, 0.654, and 0.614, with corresponding 95% CIs of 0.686–0.843, 0.510–0.798, and 0.473–0.754, in the training, internal validation, and external validation cohorts, respectively (Table 2; Fig. 2). Regression coefficients and odds ratios are reported in Supplementary Table S2.

3.3 Radiomics and DL feature selection

In the final models fitted to the full training cohort, LASSO retained 19 radiomics features: four from T2WI, four from DWI, nine from ADC, and two from contrast-enhanced images (Supplementary Table S3). Six DL features were retained: two from T2WI and four from ADC.

Grad-CAM heatmaps qualitatively illustrated regions contributing to network output. Activation appeared more central in the representative LNM-negative case and more peripheral in the LNM-positive case (Fig. 3); no quantitative spatial analysis was performed.

3.4 Comparison of the performance between radiomics model and DL model

The radiomics model achieved AUCs of 0.803 (95% CI: 0.728–0.877), 0.689 (95% CI: 0.560–0.817), and 0.665 (95% CI: 0.546–0.785) in the training, internal validation, and external validation cohorts, respectively. Corresponding DL AUCs were 0.884 (95% CI: 0.834–0.933), 0.799 (95% CI: 0.710–0.889), and 0.796 (95% CI: 0.712–0.880) (Table 2). DL AUCs were numerically higher, but differences were nonsignificant (*P*=0.086, 0.070, and 0.089, respectively; Fig. 2 and Supplementary Fig. S2).

Given the LNM class imbalance, P-R AUCs were also compared using paired bootstrap analysis. DL yielded numerically higher P-R AUCs than radiomics in all cohorts, but none of these differences was statistically significant (Supplementary Table S4).

3.5 Development and Validation of the Nomogram Model

The Clinical+Radiomics model combined the four final clinical predictors with the radiomics signature without further selection; the nomogram additionally incorporated the DL signature. Coefficients were re-estimated for each model (Table 3; Fig. 4). The final nomogram fitted to the original full training cohort achieved an AIC of 93.686. Representative LNM-negative and LNM-positive cases with nomogram-derived risks are shown in Supplementary Figs. S3 and S4, respectively.

The Clinical+Radiomics model achieved AUCs of 0.842 (95% CI: 0.777–0.907), 0.733 (95% CI: 0.591–0.875), and 0.715 (95% CI: 0.594–0.836) in the training, internal validation, and external validation cohorts, respectively. Corresponding nomogram AUCs were 0.947 (95% CI: 0.910–0.983), 0.866 (95% CI: 0.786–0.946), and 0.886 (95% CI: 0.817–0.955) (Table 2; Fig. 2).

In the training, internal validation, and external validation cohorts, respectively, the nomogram achieved significantly higher AUCs than the clinical (*P*<0.001, *P*=0.005, *P*<0.001), radiomics (*P*<0.001, *P*<0.001, *P*=0.001), and Clinical+Radiomics (*P*<0.001, *P*=0.024, *P*<0.001) models. Improvement over DL was significant in training (*P*=0.001) and external validation (*P*=0.030), but not internal validation (*P*=0.476) (Supplementary Fig. S2). The nomogram had numerically higher P-R AUCs than all comparators across cohorts; however, differences versus DL and Clinical+Radiomics were not statistically significant in internal validation (Supplementary Table S5).

Nomogram F1 scores were 0.667, 0.590, and 0.605, and accuracies were 0.854, 0.786, and 0.874 in the training, internal validation, and external validation cohorts, respectively (Table 2).

Calibration curves indicated some risk overestimation in internal validation and underestimation in external validation (Fig. 5A). In the training, internal validation, and external validation cohorts, respectively, Brier scores were 0.057, 0.110, and 0.083; calibration intercepts were −0.291, 0.028, and 0.058; and slopes were 1.176, 1.059, and 1.102. Hosmer–Lemeshow tests showed no significant lack of fit (*P*=0.443, 0.405, and 0.978, respectively).

DCA showed greater nomogram net benefit than comparators at threshold probabilities of 0.01–0.80 (training), 0.02–0.39 (internal validation), and 0.01–0.71 (external validation) (Fig. 5B–D; Supplementary Table S6), with the highest net benefit at 0.10, 0.20, and 0.30 across cohorts.

3.6 Sensitivity analysis

The prespecified sensitivity analysis included 80 patients per validation cohort with negative MR report-based LNM status; 11 internal and 9 external validation patients had pathologically confirmed LNM. The locked nomogram used the training OOF-derived threshold of 0.161 without refitting, recalibration, feature selection, or threshold optimization.

In the internal and external validation subgroups, respectively, AUCs were 0.843 (95% CI: 0.717–0.969) and 0.858 (95% CI: 0.756–0.959); sensitivities, 0.636 and 0.444; specificities, 0.826 and 0.930; negative predictive values (NPVs), 0.934 and 0.930; accuracies, 0.800 and 0.875; P-R AUCs, 0.535 (95% bootstrap CI: 0.272–0.814) and 0.497 (95% bootstrap CI: 0.170–0.778); and F1 scores, 0.467 and 0.444 (Supplementary Table S7).

4. Discussion

In this two-center study, a nomogram integrating clinical variables, multisequence MRI radiomics, and deep transfer-learning features showed good discrimination for preoperative LNM prediction in EC in held-out internal and external validation cohorts. Adding the DL signature significantly improved discrimination over Clinical+Radiomics across all cohorts, suggesting complementary predictive information. Lower validation than training performance nevertheless indicates potential optimism and supports prospective multicenter validation.

Clinical predictors included HE4, MR report-based MTD, MR report-based DMI, and MR report-based LNM status. HE4 is associated with tumor progression and adverse outcomes[29], while tumor size and DMI indicate aggressive EC[30]. Limited HE4 performance in early or low-volume disease and MRI reliance on reader experience and morphological criteria may explain the clinical model's modest discrimination[31].

Most retained radiomics features originated from DWI or ADC, suggesting a contribution from diffusion information, complemented by anatomical and perfusion information from T2WI and contrast-enhanced imaging[32, 33]. Modest validation performance suggests that handcrafted features may incompletely capture nodal metastatic phenotypes; differences in populations, acquisition, segmentation, and validation may explain lower performance than previous reports[34].

DL learns multilevel representations beyond predefined descriptors[35, 36], and previous studies support its feasibility for MRI-based LNM prediction and integration with radiomics and clinical variables in EC[37, 38]. Here, DL AUCs exceeded radiomics AUCs numerically but not significantly in any cohort. Its significant incremental value within Clinical+Radiomics suggests complementary imaging information.

Grad-CAM coarsely localizes regions contributing to predictions[28]. Activation appeared more central in the representative LNM-negative case and more peripheral in the LNM-positive case. Given the limited accuracy,

reproducibility, and robustness of saliency maps[39, 40], these exploratory observations require cohort-level analysis and pathological correlation before biological interpretation.

In patients with negative MR report-based LNM status, the nomogram retained discrimination in both validation cohorts. Size- and morphology-based MRI assessment can miss normal-sized metastatic nodes or misclassify reactive enlargement[41, 42]. However, limited sensitivity leaves residual occult metastatic risk after a negative prediction, so the nomogram should not justify omitting pathological nodal staging.

The retrospective design may introduce selection bias. In addition, the extent of pathological nodal assessment varied, and para-aortic nodes were not sampled in all patients; occult metastases in unsampled regions may therefore have been missed. The relatively small number of LNM-positive events and high-dimensional multistep modeling further increase the risk of overfitting. ResNet-50 tuning and fine-tuning preceded downstream cross-validation and were not repeated within folds; OOF estimates for DL-containing models may therefore be optimistic and should not be treated as unbiased validation of the entire pipeline. Held-out validation, especially external validation, provides the primary evidence of generalizability. Generalizability is further limited by two centers, restricted scanners and protocols, and exclusion of 84 patients with small or MRI-occult tumors, which may introduce spectrum bias. Manual multislice segmentation is time-consuming and operator-dependent. Incremental discrimination does not establish improved clinical decisions or patient outcomes. Prospective multicenter studies should validate a locked pipeline with automated segmentation and prespecified clinical thresholds. The nomogram remains complementary to pathological staging and multidisciplinary judgment.

5. Conclusion

The integrated nomogram showed good discrimination for LNM in both validation cohorts, including patients with negative MR report-based LNM status. Adding the DL signature significantly improved discrimination over Clinical+Radiomics across all three cohorts. Prospective multicenter validation is required before clinical implementation as a complementary risk-stratification tool.

Declarations

Author Contribution

T.Z. contributed to conceptualization, methodology, and writing the original draft. F.H. and D.D. contributed to investigation and data curation. X.L. and J.D. contributed to data curation. Y.T. and J.P. contributed to investigation and resources. L.Y. contributed to validation and supervision. Q.S. contributed to validation and manuscript review and editing. D.L. contributed to supervision, project administration, and manuscript review and editing. All authors contributed to the interpretation of the results, critically reviewed the manuscript, and approved the final version for publication.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available because they contain potentially sensitive patient information but are available from the corresponding author on reasonable request, subject to institutional and ethical approval.

References

1. Crosbie EJ, Kitson SJ, Mcalpine JN, et al. Endometrial cancer [J]. Lancet, 2022, 399(10333): 1412–1428.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020 [J]. CA Cancer J Clin, 2020, 70(1): 7–30.
3. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015 [J]. CA Cancer J Clin, 2016, 66(2): 115–132.
4. Lukanović D, Matjašič M, Kobal B. Accuracy of preoperative sampling diagnosis for predicting final pathology in patients with endometrial carcinoma: a review [J]. Transl Cancer Res, 2020, 9(12): 7785–7796.
5. Abu-Rustum N, Yashar C, Arend R, et al. Uterine Neoplasms, Version 1.2023, NCCN Clinical Practice Guidelines in Oncology [J]. J Natl Compr Canc Netw, 2023, 21(2): 181–209.
6. Bogani G, Casarin J, Leone Roberti Maggiore U, et al. Survival outcomes in endometrial cancer patients having lymphadenectomy, sentinel node mapping followed by lymphadectomy and sentinel node mapping alone: Long-term results of a propensity-matched analysis [J]. Gynecol Oncol, 2020, 158(1): 77–83.
7. Bogani G, Di Donato V, Papadia A, et al. Evaluating long-term outcomes of three approaches to retroperitoneal staging in endometrial cancer [J]. Gynecol Oncol, 2022, 166(2): 277–283.
8. Benedetti Panici P, Basile S, Maneschi F, et al. Systematic pelvic lymphadenectomy vs. no lymphadenectomy in early-stage endometrial carcinoma: randomized clinical trial [J]. J Natl Cancer Inst, 2008, 100(23): 1707–1716.
9. Kitchener H, Swart AM, Qian Q, et al. Efficacy of systematic pelvic lymphadenectomy in endometrial cancer (MRC ASTEC trial): a randomised study [J]. Lancet, 2009, 373(9658): 125–136.
10. Proppe L, Alkatout I, Koch R, et al. Impact of lymphadenectomy on short- and long-term complications in patients with endometrial cancer [J]. Arch Gynecol Obstet, 2022, 306(3): 811–819.
11. Haldorsen IS, Salvesen HB. What Is the Best Preoperative Imaging for Endometrial Cancer? [J]. Curr Oncol Rep, 2016, 18(4): 25.
12. Cè M, Chiriac MD, Cozzi A, et al. Decoding Radiomics: A Step-by-Step Guide to Machine Learning Workflow in Hand-Crafted and Deep Learning Radiomics Studies [J]. Diagnostics (Basel), 2024, 14(22).
13. Kido A, Himoto Y, Kurata Y, et al. Preoperative Imaging Evaluation of Endometrial Cancer in FIGO 2023 [J]. J Magn Reson Imaging, 2024, 60(4): 1225–1242.
14. Zhang Z, Wan X, Lei X, et al. Intra- and peri-tumoral MRI radiomics features for preoperative lymph node metastasis prediction in early-stage cervical cancer [J]. Insights Imaging, 2023, 14(1): 65.
15. He Y, Duan S, Wang W, et al. Integrative radiomics clustering analysis to decipher breast cancer heterogeneity and prognostic indicators through multiparametric MRI [J]. NPJ Breast Cancer, 2024, 10(1): 72.
16. Nayak SS, Pendem S, Menon GR, et al. Quality Assessment of MRI-Radiomics-Based Machine Learning Methods in Classification of Brain Tumors: Systematic Review [J]. Diagnostics (Basel), 2024, 14(23).

17. Valizadeh P, Jannatdoust P, Pahlevan-Fallahy MT, et al. Diagnostic accuracy of radiomics and artificial intelligence models in diagnosing lymph node metastasis in head and neck cancers: a systematic review and meta-analysis [J]. Neuroradiology, 2024.
18. Xiong L, Chen C, Lin Y, et al. A computer-aided determining method for the myometrial infiltration depth of early endometrial cancer on MRI images [J]. Biomed Eng Online, 2023, 22(1): 103.
19. Dong HC, Dong HK, Yu MH, et al. Using Deep Learning with Convolutional Neural Network Approach to Identify the Invasion Depth of Endometrial Cancer in Myometrium Using MR Images: A Pilot Study [J]. Int J Environ Res Public Health, 2020, 17(16).
20. Toseef M, Olayemi Petinrin O, Wang F, et al. Deep transfer learning for clinical decision-making based on high-throughput data: comprehensive survey with benchmark results [J]. Brief Bioinform, 2023, 24(4).
21. Saida T, Gu W, Hoshiai S, et al. Artificial Intelligence in Obstetric and Gynecological MR Imaging [J]. Magn Reson Med Sci, 2024.
22. Qi X. Artificial intelligence-assisted magnetic resonance imaging technology in the differential diagnosis and prognosis prediction of endometrial cancer [J]. Sci Rep, 2024, 14(1): 26878.
23. Concin N, Matias-Guiu X, Cibula D, et al. ESGO-ESTRO-ESP guidelines for the management of patients with endometrial carcinoma: update 2025 [J]. Lancet Oncol, 2025, 26(8): e423-e435.
24. Chiang S, Tessier-Cloutier B, Klein E, et al. Establishing guidelines for sentinel lymph node ultrastaging in endometrial cancer [J]. Int J Gynecol Cancer, 2024, 34(5): 681–688.
25. Berek JS, Matias-Guiu X, Creutzberg C, et al. FIGO staging of endometrial cancer: 2023 [J]. Int J Gynaecol Obstet, 2023, 162(2): 383–394.
26. Iitsuka C, Asami Y, Hirose Y, et al. Preoperative Magnetic Resonance Imaging versus Intraoperative Frozen Section Diagnosis for Predicting the Deep Myometrial Invasion in Endometrial Cancer: Our Experience and Literature Review [J]. J Obstet Gynaecol Res, 2021, 47(9): 3331–3338.
27. Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning [M]//Huang D S, Zhang X P, Huang G B. Advances in Intelligent Computing, Pt 1, Proceedings. 2005: 878–887.
28. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization; proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, ITALY, F Oct 22–29, 2017 [C]. 2017.
29. Lan T, Mu C, Wang Z, et al. Diagnostic and Prognostic Values of Serum EpCAM, TGM2, and HE4 Levels in Endometrial Cancer [J]. Front Oncol, 2020, 10: 1697.
30. Koskas M, Chereau E, Ballester M, et al. Accuracy of a nomogram for prediction of lymph-node metastasis detected with conventional histopathology and ultrastaging in endometrial cancer [J]. Br J Cancer, 2013, 108(6): 1267–1272.
31. Nougaret S, Horta M, Sala E, et al. Endometrial Cancer MRI staging: Updated Guidelines of the European Society of Urogenital Radiology [J]. Eur Radiol, 2019, 29(2): 792–805.
32. Ou R, Peng Y. Preoperative risk stratification of early-stage endometrial cancer assessed by multimodal magnetic resonance functional imaging [J]. Magn Reson Imaging, 2024, 117: 110283.
33. Yoshida M, Saida T, Shibuki S, et al. The Utility of Apparent Water Diffusion Coefficient Maps for Evaluating the Presence of Myometrial Invasion in Patients with Endometrial Cancer [J]. Magn Reson Med Sci, 2024.

34. Yan BC, Li Y, Ma FH, et al. Radiologists with MRI-based radiomics aids to predict the pelvic lymph node metastasis in endometrial cancer: a multicenter study [J]. Eur Radiol, 2021, 31(1): 411–422.
35. He KM, Zhang XY, Ren SQ, et al. Deep Residual Learning for Image Recognition; proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, F Jun 27–30, 2016 [C]. 2016.
36. Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks? [J]. Advances in neural information processing systems, 2014, 27.
37. Wang Y, Liu W, Lu Y, et al. Fully Automated Identification of Lymph Node Metastases and Lymphovascular Invasion in Endometrial Cancer From Multi-Parametric MRI by Deep Learning [J]. J Magn Reson Imaging, 2024, 60(6): 2730–2742.
38. Yang J, Cao Y, Zhou F, et al. Combined deep-learning MRI-based radiomic models for preoperative risk classification of endometrial endometrioid adenocarcinoma [J]. Front Oncol, 2023, 13: 1231497.
39. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging [J]. Radiology: Artificial Intelligence, 2021, 3(6): e200267.
40. Zhang J, Chao H, Dasegowda G, et al. Revisiting the trustworthiness of saliency methods in radiology AI [J]. Radiology: Artificial Intelligence, 2023, 6(1): e220221.
41. Lai G, Rockall AG. Lymph node imaging in gynecologic malignancy [J]. Semin Ultrasound CT MR, 2010, 31(5): 363–376.
42. Thoeny HC, Froehlich JM, Triantafyllou M, et al. Metastases in normal-sized pelvic lymph nodes: detection with diffusion-weighted MR imaging [J]. Radiology, 2014, 273(1): 125–135.

Tables

Table 1 Clinical characteristics of patients in different sets

Characteristic	Training Set	Internal Validation Set	External Validation Set	P
Sample size	268	117	135	
Age (years, mean ± SD)	54.6 ± 10.7	55.4 ± 9.2	54.2 ± 9.7	0.635
BMI (kg/m², mean ± SD)	25.5 ± 4.8	25.9 ± 5.5	26.9 ± 5.1	0.032
Postmenopausal status, n (%)				0.218
No	52 (19.4%)	30 (25.6%)	35 (25.9%)	
Yes	216 (80.6%)	87 (74.4%)	100 (74.1%)	
Reproductive history, n (%)				0.024
No	58 (21.6%)	23 (19.7%)	44 (32.6%)	
Yes	210 (78.4%)	94 (80.3%)	91 (67.4%)	
CA125, n (%)				0.769
Normal	95 (35.4%)	43 (36.8%)	44 (32.6%)	
Abnormal	173 (64.6%)	74 (63.2%)	91 (67.4%)	
HE4, n (%)				0.600
Normal	162 (60.4%)	71 (60.7%)	75 (55.6%)	
Abnormal	106 (39.6%)	46 (39.3%)	60 (44.4%)	
MR report-based DMI, n (%)				<0.001
No	220 (82.1%)	94 (80.3%)	75 (55.6%)	
Yes	48 (17.9%)	23 (19.7%)	60 (44.4%)	
MR report-based MTD (cm, mean ± SD)	4.6 ± 1.8	4.2 ± 1.5	4.4 ± 1.7	0.105
MR report-based LNM status, n (%)				0.190
No	158 (59.0%)	80 (68.4%)	80 (59.3%)	
Yes	110 (41.0%)	37 (31.6%)	55 (40.7%)	
Histological type and grade, n (%)				0.450
G1+G2	196 (73.1%)	90 (76.9%)	106 (78.5%)	

Characteristic	Training Set	Internal Validation Set	External Validation Set	P
G3+non-endometrioid carcinoma	72 (26.9%)	27 (23.1%)	29 (21.5%)	
Pathological DMI, n (%)				0.388
No	180 (67.2%)	76 (65.0%)	98 (72.6%)	
Yes	88 (32.8%)	41 (35.0%)	37 (27.4%)	
LVSI, n (%)				0.129
No	204 (76.1%)	85 (72.6%)	112 (83.0%)	
Yes	64 (23.9%)	32 (27.4%)	23 (17.0%)	
FIGO stage, n (%)				0.372
I	165 (61.6%)	71 (60.7%)	82 (60.7%)	
II	53 (19.8%)	21 (17.9%)	20 (14.8%)	
III	25 (9.3%)	9 (7.7%)	20 (14.8%)	
IV	25 (9.3%)	16 (13.7%)	13 (9.6%)	
Pathological LNM, n (%)				0.675
No	227 (84.7%)	95 (81.2%)	114 (84.4%)	
Yes	41 (15.3%)	22 (18.8%)	21 (15.6%)	

Note: Continuous variables were compared by one-way analysis of variance and categorical variables by Pearson's chi-square test. FIGO stage was retrospectively harmonized according to the 2023 FIGO staging system for all patients. BMI, body mass index; CA125, cancer antigen 125; HE4, human epididymis protein 4; MR, magnetic resonance; DMI, deep myometrial invasion; MTD, maximum tumor diameter; LVSI, lymphovascular space invasion; FIGO, International Federation of Gynecology and Obstetrics; LNM, lymph node metastasis; SD, standard deviation.

Table 2. Performance of the models

Set	Model	ROC AUC (95% CI)	Acc	Sen	Spe	Pre	P-R AUC (95% CI)	F1
Training set (OOF)	Clinical	0.765 (0.686–0.843)	0.683	0.854	0.652	0.307	0.397 (0.272–0.546)	0.452
	Radiomics	0.803 (0.728–0.877)	0.780	0.707	0.793	0.382	0.506 (0.359–0.651)	0.496
	DL	0.884 (0.834–0.933)	0.828	0.805	0.833	0.465	0.658 (0.511–0.780)	0.589
	Clinical+Radiomics	0.842 (0.777–0.907)	0.757	0.780	0.753	0.364	0.582 (0.422–0.720)	0.496
	Nomogram	0.947 (0.910–0.983)	0.854	0.951	0.837	0.513	0.838 (0.735–0.918)	0.667
Internal validation set	Clinical	0.654 (0.510–0.798)	0.692	0.545	0.726	0.316	0.331 (0.203–0.541)	0.400
	Radiomics	0.689 (0.560–0.817)	0.778	0.318	0.884	0.389	0.372 (0.227–0.578)	0.350
	DL	0.799 (0.710–0.889)	0.607	0.818	0.558	0.300	0.489 (0.318–0.685)	0.439
	Clinical+Radiomics	0.733 (0.591–0.875)	0.838	0.500	0.916	0.579	0.476 (0.295–0.706)	0.537
	Nomogram	0.866 (0.786–0.946)	0.786	0.818	0.779	0.462	0.599 (0.410–0.799)	0.590
External validation set	Clinical	0.614 (0.473–0.754)	0.504	0.762	0.456	0.205	0.256 (0.144–0.452)	0.323
	Radiomics	0.665 (0.546–0.785)	0.681	0.476	0.719	0.238	0.341 (0.179–0.535)	0.317
	DL	0.796 (0.712–0.880)	0.807	0.524	0.860	0.407	0.350 (0.220–0.565)	0.458
	Clinical+Radiomics	0.715 (0.594–0.836)	0.681	0.667	0.684	0.280	0.358 (0.192–0.556)	0.394
	Nomogram	0.886 (0.817–0.955)	0.874	0.619	0.921	0.591	0.675 (0.477–0.837)	0.605

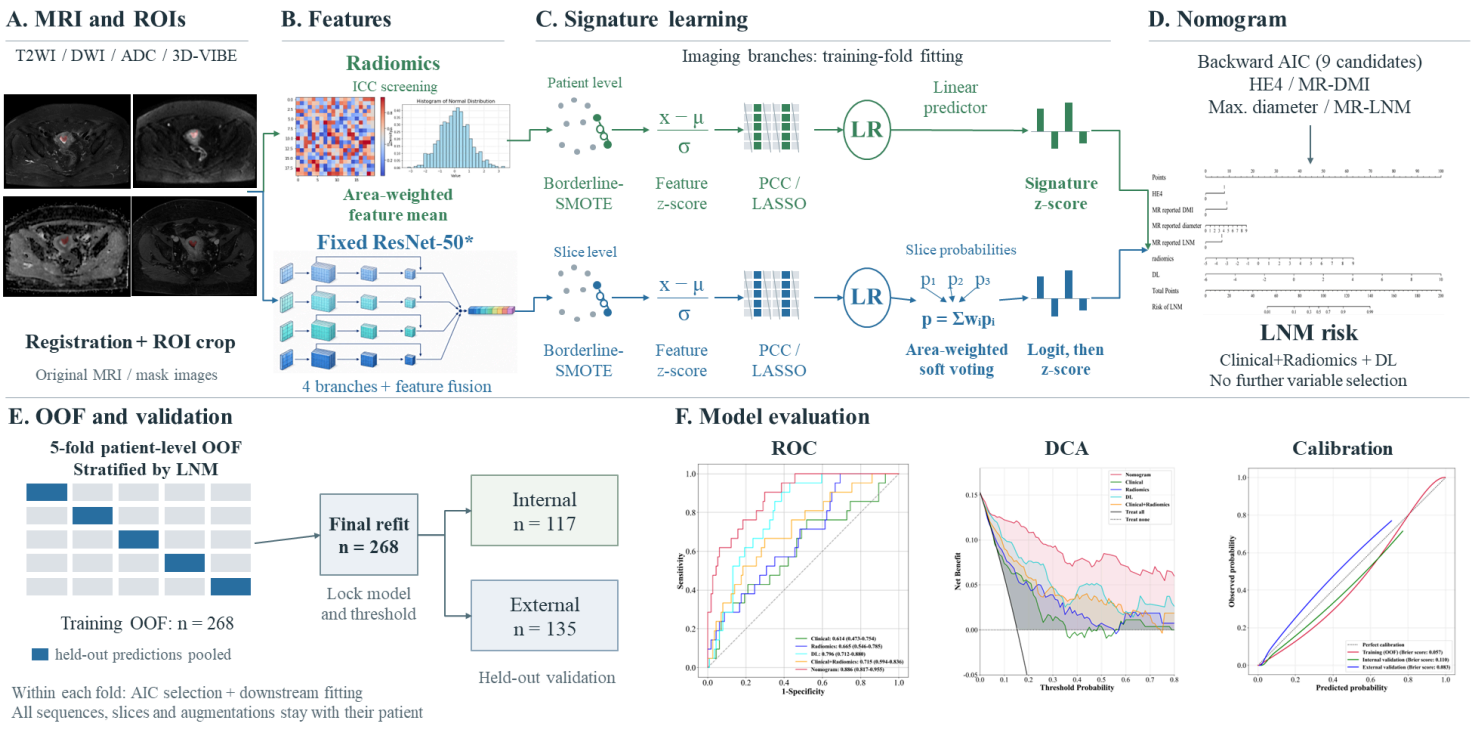
Note: Training estimates were calculated from five-fold OOF predictions for the original training records; no metric was calculated on SMOTE-generated observations. For each model, the threshold maximizing the Youden index in training OOF predictions was locked and applied unchanged to both validation cohorts. ROC AUC CIs used the DeLong method; P-R AUC CIs used 1000 patient-level bootstrap replicates (seed 978). Acc, accuracy; Sen, sensitivity; Spe, specificity; Pre, precision (positive predictive value); F1, F1 score; DL, deep learning; OOF, out of fold; ROC AUC, area under the receiver operating characteristic curve; P-R AUC, area under the precision–recall curve; SMOTE, Synthetic Minority Oversampling Technique; CI, confidence interval. The locked nomogram classification threshold was 0.161.

Table 3. Variables and coefficients of the deep learning radiomics nomogram

Variable	β	Adjusted OR (95% CI)	P
Intercept	-7.876	Not applicable	<0.001
HE4 (abnormal vs. normal)	1.372	3.945 (1.098, 14.167)	0.035
MR report-based DMI (positive vs. negative)	1.639	5.149 (1.287, 20.606)	0.021
MR report-based maximum tumor diameter (per 1-cm increase)	0.298	1.347 (0.955, 1.898)	0.089
MR report-based LNM status (positive vs. negative)	1.637	5.142 (1.549, 17.064)	0.007
Radiomics signature (per 1-SD increase)	1.662	5.267 (2.502, 11.088)	<0.001
DL signature (per 1-SD increase)	1.941	6.964 (3.523, 13.768)	<0.001

Note: The integrated logistic model was fitted to the original training records without SMOTE. CIs and P values are Wald estimates from the same model used to obtain predictions. HE4, human epididymis protein 4; MR, magnetic resonance; DMI, deep myometrial invasion; LNM, lymph node metastasis; DL, deep learning; OR, odds ratio; CI, confidence interval; SD, standard deviation; SMOTE, Synthetic Minority Oversampling Technique.

Figures



* ResNet tuning, fine-tuning and early stopping preceded downstream OOF. DL-containing OOF estimates may remain optimistic. Held-out validation provides the primary evidence of generalizability.

Figure 1

Workflow for developing and evaluating the multisequence MRI-based nomogram for preoperative prediction of lymph node metastasis in endometrial cancer. (A) MRI registration and tumor-mask-based region-of-interest cropping. (B) Extraction of handcrafted radiomics features and deep features using a fixed ResNet-50 with four sequence-specific branches and feature fusion. (C) Separate radiomics and deep-learning signature development using Borderline-SMOTE, input-feature z-score standardization, PCC/LASSO feature selection, and logistic regression. For radiomics, slice-level features were first aggregated to the patient level using tumor-area-weighted averaging before downstream signature development. For DL, slice-level probabilities were aggregated within each patient using tumor-area-weighted soft voting and then logit-transformed to obtain the patient-level DL signature. The resulting patient-level radiomics and DL signatures were subsequently z-score standardized separately using training-derived parameters. (D) Integration of the standardized signatures with clinical predictors into the nomogram. (E) Patient-level five-fold downstream out-of-fold (OOF) evaluation in the training cohort (n = 268), followed by downstream refitting on the entire training cohort and locked evaluation in the internal (n = 117) and external (n = 135) validation cohorts. ResNet-50 fine-tuning, hyperparameter tuning, and early stopping occurred outside the OOF procedure. (F) Representative receiver operating characteristic (ROC), decision curve analysis (DCA), and calibration plots. Feature graphics are schematic; the ResNet-50 illustration is AI-generated.

MRI, magnetic resonance imaging; ResNet-50, 50-layer residual network; Borderline-SMOTE, Borderline Synthetic Minority Oversampling Technique; AI, artificial intelligence; PCC, Pearson correlation coefficient; LASSO, least absolute shrinkage and selection operator; DL, deep learning; LNM, lymph node metastasis.

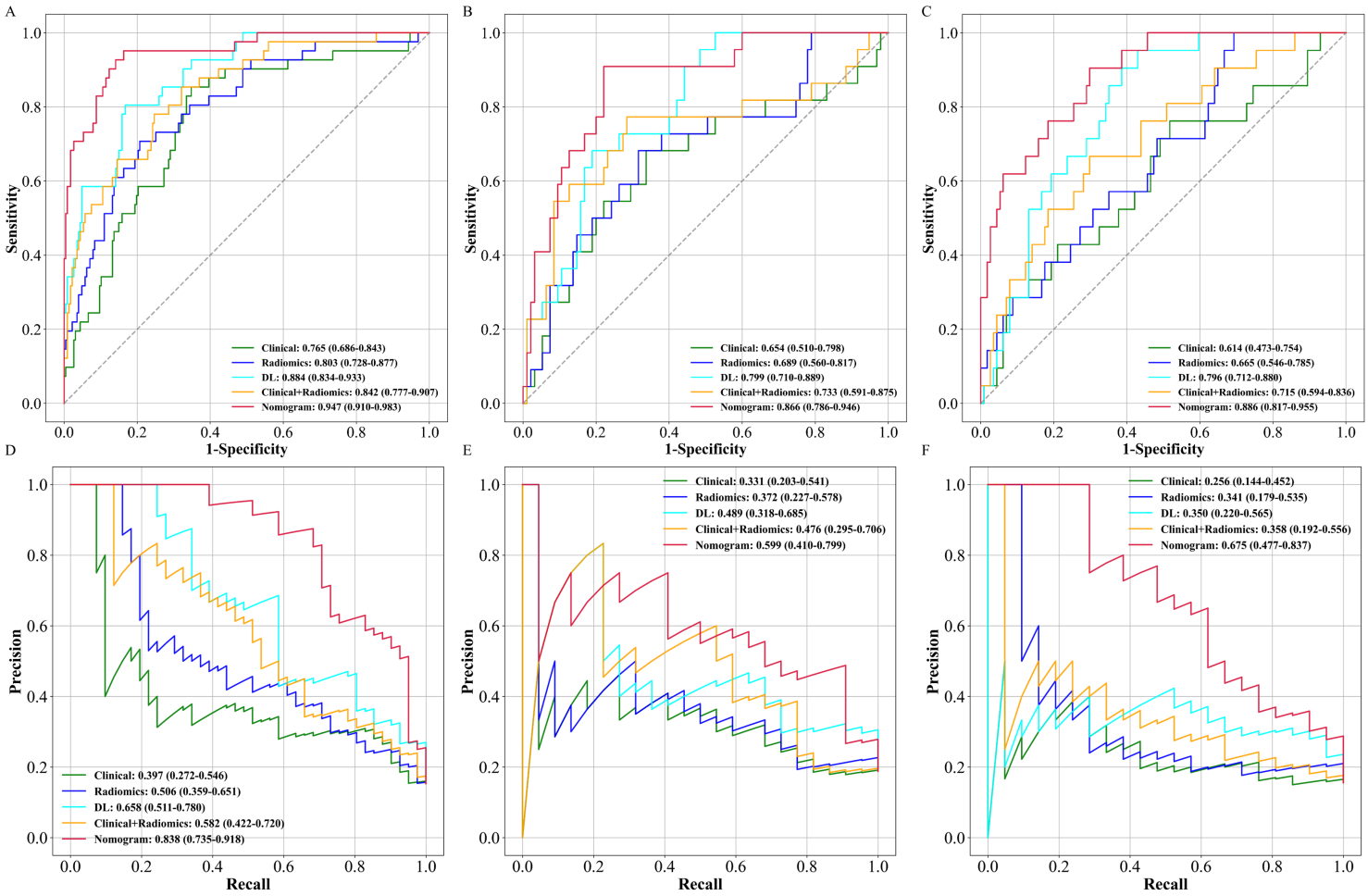


Figure 2

Performance of five predictive models. (A–C) Receiver operating characteristic (ROC) curves of the five predictive models in (A) the training cohort based on out-of-fold (OOF) predictions, (B) the internal validation cohort, and (C) the external validation cohort. (D–F) Precision–recall (P-R) curves of the predictive models in (D) the training cohort based on OOF predictions, (E) the internal validation cohort, and (F) the external validation cohort. AUC, area under the curve; DL, deep learning.

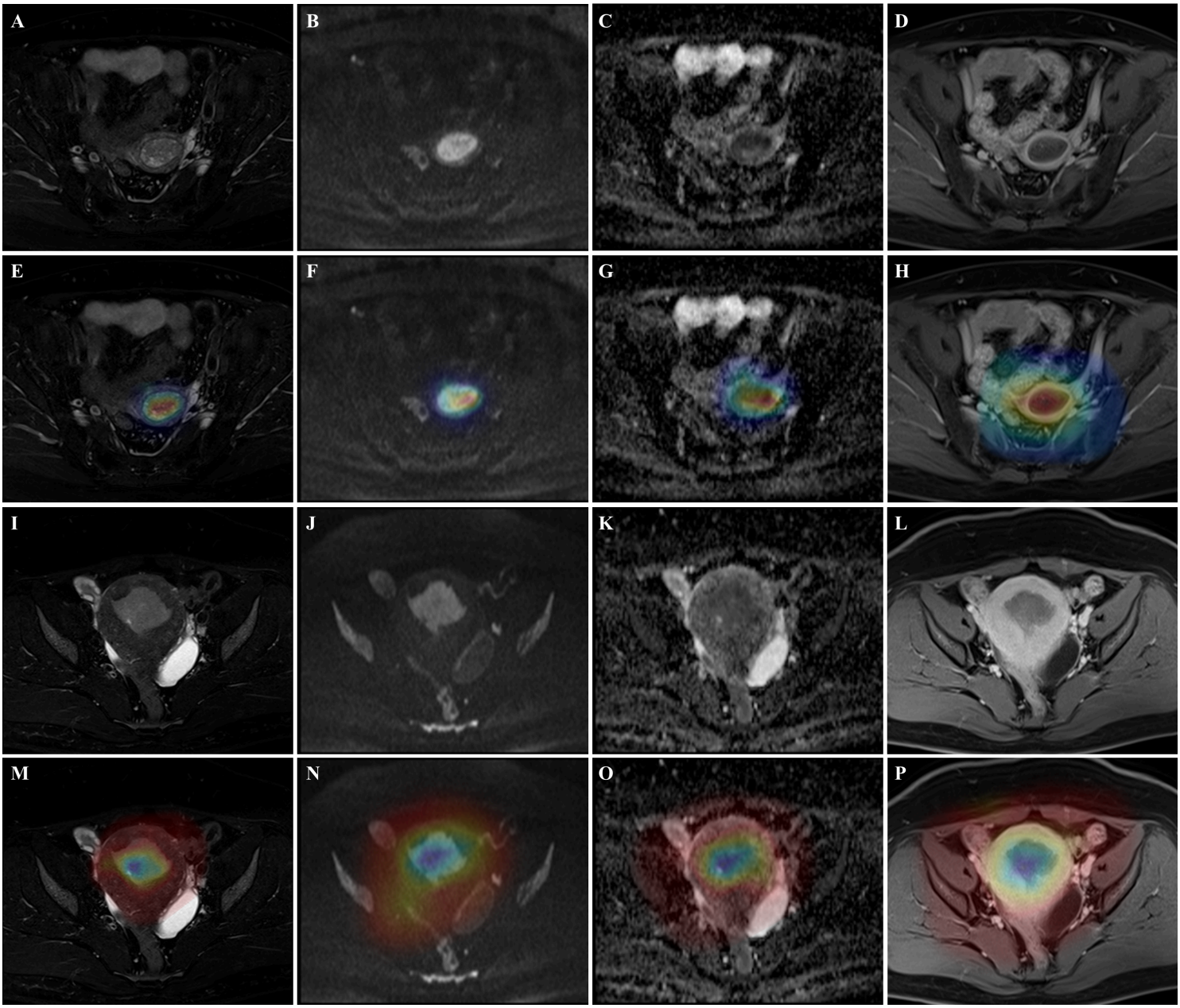


Figure 3

Representative multi-sequence MR images and Grad-CAM visualizations of LNM-negative and LNM-positive endometrial cancer. Panels A–H show a 61-year-old patient with pathologically confirmed LNM-negative endometrial cancer: A–D correspond to T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) map, and three-dimensional volumetric interpolated breath-hold examination (3D-VIBE), respectively; E–H show the corresponding gradient-weighted class activation mapping (Grad-CAM) overlays. Panels I–P show a 72-year-old patient with pathologically confirmed LNM-positive endometrial cancer: I–L correspond to T2WI, DWI, ADC map, and 3D-VIBE, respectively; M–P show the corresponding Grad-CAM overlays. All images were obtained from the axial slice containing the largest tumor cross-section. The Grad-CAM maps provide a qualitative visualization of image regions contributing to the convolutional neural network (CNN) output. MR, magnetic resonance; LNM, lymph node metastasis.

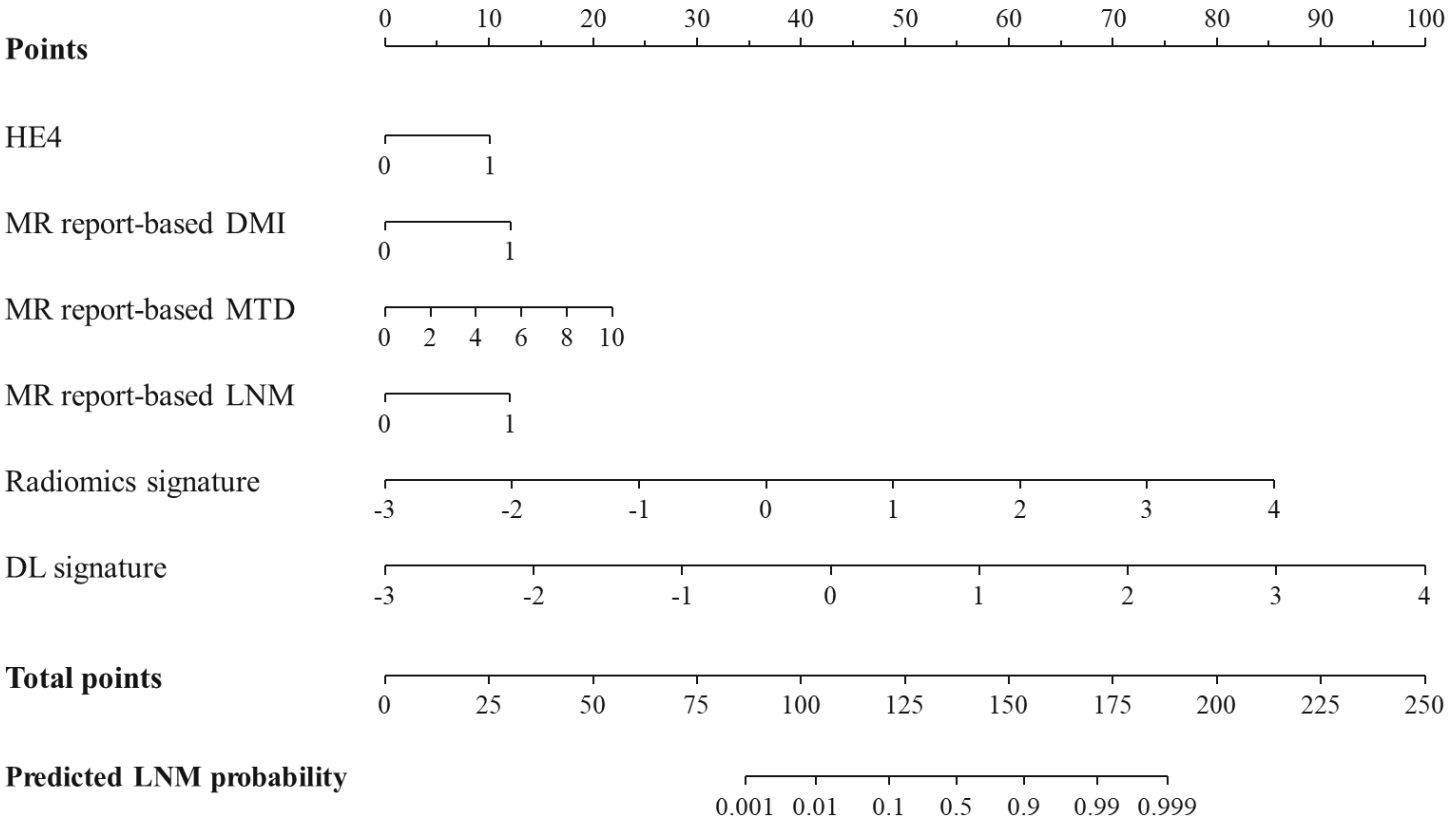


Figure 4

Nomogram for LNM risk prediction in EC patients. EC: endometrial cancer; HE4: human epididymis protein 4; MR: magnetic resonance; DMI: deep myometrial invasion; MTD: maximum tumor diameter; LNM: lymph node metastasis; DL: deep learning

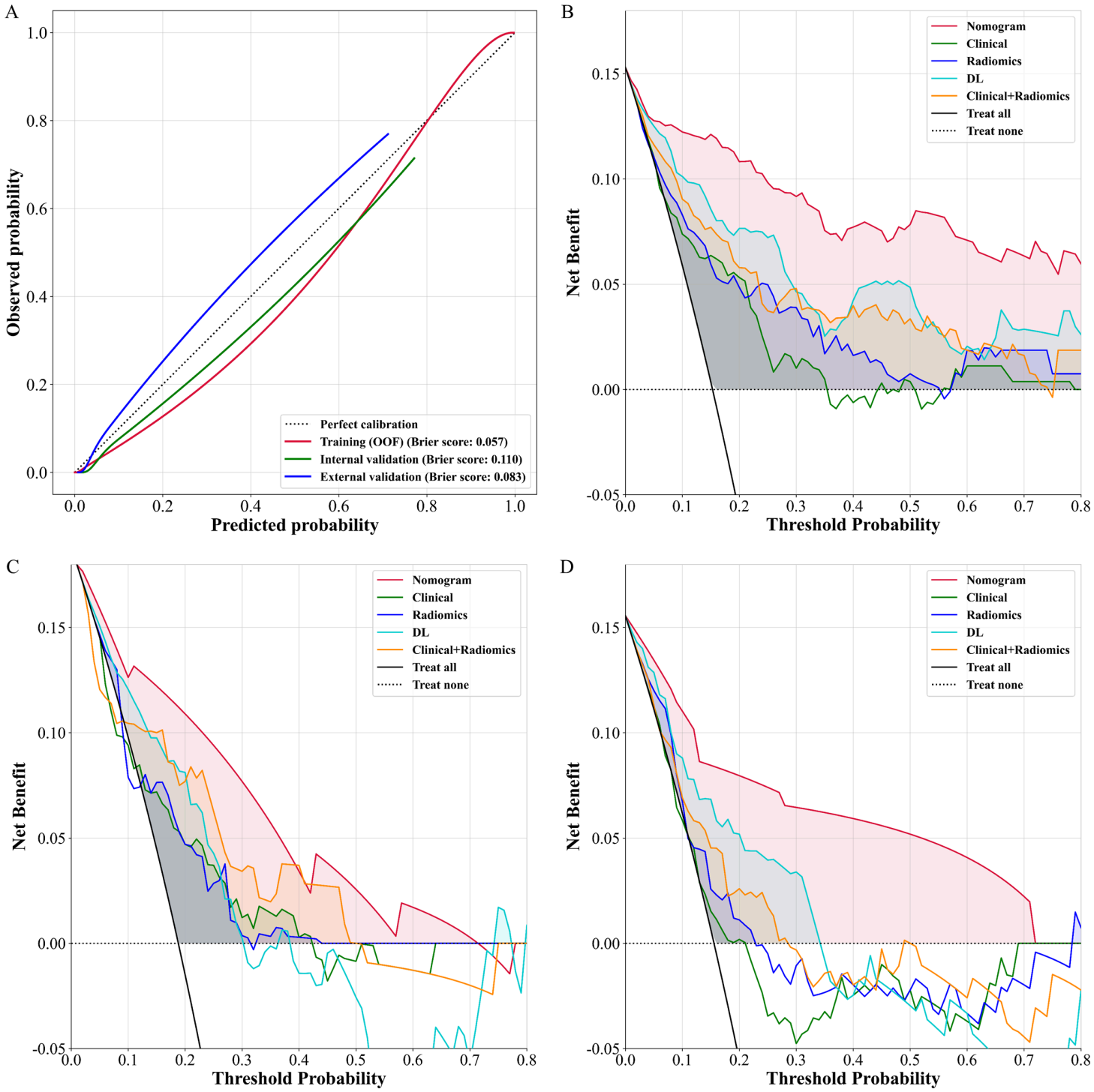


Figure 5

Calibration curves and decision curve analysis (DCA). (A) Calibration curves of the nomogram in the training cohort based on out-of-fold (OOF) predictions, internal validation cohort, and external validation cohort. Cubic B-spline calibration curves were fitted separately within each cohort over its observed probability range. The 45° reference line represents perfect calibration. The Brier scores for each cohort are shown in the legend. (B–D) Decision curve analyses comparing the net benefit of the nomogram, clinical model, radiomics model, deep-learning (DL) model, and Clinical+Radiomics model in (B) the training cohort based on OOF predictions, (C) the

internal validation cohort, and (D) the external validation cohort. The black line represents the strategy of treating all patients, and the black dotted line represents the strategy of treating none.

Supplementary Files

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

- Supplementaryrevised20260908.docx