BMC Medical Imaging

RESEARCH

Open Access

Ultrasound-based machine learning model to predict the risk of endometrial cancer among postmenopausal women

Yi-Xin Li[1†], Yu Lu[1†], Zhe-Ming Song[2], Yu-Ting Shen[1], Wen Lu[1] and Min Ren[1,3*]

Abstract

Background Current ultrasound-based screening for endometrial cancer (EC) primarily relies on endometrial thickness (ET) and morphological evaluation, which suffer from low specificity and high interobserver variability. This study aimed to develop and validate an artificial intelligence (AI)-driven diagnostic model to improve diagnostic accuracy and reduce variability.

Methods A total of 1,861 consecutive postmenopausal women were enrolled from two centers between April 2021 and April 2024. Super-resolution (SR) technique was applied to enhance image quality before feature extraction. Radiomics features were extracted using Pyradiomics, and deep learning features were derived from convolutional neural network (CNN). Three models were developed: (1) R model: radiomics-based machine learning (ML) algorithms; (2) CNN model: image-based CNN algorithms; (3) DLR model: a hybrid model combining radiomics and deep learning features with ML algorithms.

Results Using endometrium-level regions of interest (ROI), the DLR model achieved the best diagnostic performance, with an area under the receiver operating characteristic curve (AUROC) of 0.893 (95% CI: 0.847–0.932), sensitivity of 0.847 (95% CI: 0.692–0.944), and specificity of 0.810 (95% CI: 0.717–0.910) in the internal testing dataset. Consistent performance was observed in the external testing dataset (AUROC 0.871, sensitivity 0.792, specificity 0.829). The DLR model consistently outperformed both the R and CNN models. Moreover, endometrium-level ROIs yielded better results than uterine-corpus-level ROIs.

Conclusions This study demonstrates the feasibility and clinical value of AI-enhanced ultrasound analysis for EC detection. By integrating radiomics and deep learning features with SR-based image preprocessing, our model improves diagnostic specificity, reduces false positives, and mitigates operator-dependent variability. This non-invasive approach offers a more accurate and reliable tool for EC screening in postmenopausal women.

Clinical trial number Not applicable.

Keywords Ultrasound, Machine learning, Convolutional neural networks, Deep learning, Radiomics, Endometrial cancer

[†]Yi-Xin Li and Yu Lu contributed equally to this work.

*Correspondence:
Min Ren
renmin20060803@126.com

Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Endometrial cancer (EC) is the most common malignancy of the female genital tract in middle- and high-income countries, with its incidence increasing by 132% over the past 30 years [1]. While early-stage EC is often curable, delayed diagnosis significantly worsens outcomes [2]. Ultrasound is the first-line imaging modality for EC risk assessment, but current methods, particularly endometrial thickness (ET) measurement, suffer from low specificity and lead to unnecessary invasive procedures [3]. More advanced techniques like Doppler imaging and morphological evaluation are also limited by operator dependency and inconsistent performance [4, 5]. These limitations underscore the urgent need for a more objective, automated, and generalizable method for early EC risk stratification.

In recent years, artificial intelligence (AI) has shown promising results in medical imaging. Convolutional neural networks (CNNs), vision transformers (ViTs), and their integration have become crucial for both classification and segmentation tasks [6, 7]. For example, AI-based models have achieved excellent performance in detecting lung cancer [8], skin cancer diagnosis [9], and identifying lymph node metastasis in breast cancer [10]. In gynecologic oncology, AI has also been applied to ovarian cancer diagnosis [11]. Radiomics enables high-throughput extraction of quantitative imaging features, capturing subtle textural and intensity-based characteristics from ultrasound images that may not be visually discernible [12]. CNN, in contrast, autonomously learn hierarchical representations directly from raw images, capturing spatial dependencies and complex structural patterns without requiring manual feature design [13]. While radiomics provides handcrafted, interpretable imaging descriptors, it relies on predefined feature sets and may miss non-linear or abstract patterns. CNNs, on the other hand, excel at extracting deep, high-level semantic features but often lack interpretability. Prior studies have shown that the integration of radiomics and CNN-based features improves classification accuracy in various oncological imaging tasks compared to either method alone [14, 15]. However, the use of deep learning radiomics to predict endometrial cancer risk has not yet been reported.

A major challenge in developing ultrasound-based AI models is that ultrasound images often suffer from low resolution, noise, and artifacts, which can degrade the quality of extracted radiomic and deep learning features [16]. To address these limitations, super-resolution (SR) techniques have been introduced to enhance image quality by reconstructing high-resolution images from low-resolution inputs [17]. SR can reduce noise, improve structural details, and enhance contrast, thereby facilitating more reliable feature extraction and increasing the robustness of AI-based diagnostic models [17].

In this study, we aimed to develop and validate a deep learning radiomics (DLR) AI-based model to predict EC risk. By leveraging automated DLR features and AI algorithms, our model reduces operator dependency, improves reproducibility, and ensures consistent performance across different imaging settings. We also applied a standardized image preprocessing pipeline to enhance image consistency and stability, and collected data from two independent hospitals to improve model generalizability.

Methods

This was an observational study conducted at two centers, with data collected both retrospectively and prospectively. The study population included consecutive postmenopausal women who underwent both gynecologic ultrasound and endometrial pathology examinations at two hospitals. The study adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [18]. Patients retrospectively collected from hospital one between April 2021 and June 2023 were used as the training set for model development and sensitivity analyses. Patients prospectively recruited from hospital one between July 2023 and April 2024 constituted the internal testing set, used to evaluate the model's performance in a temporally distinct yet institutionally consistent cohort. Patients retrospectively collected from hospital two between January 2019 and December 2023 comprised the external testing set, serving to assess generalizability across institutions. This dataset partitioning strategy, based on both temporal and institutional separation, ensures independence among datasets, minimizes the risk of data leakage, and enhances the external validity of the model. The inclusion of a prospective internal testing set al.so strengthens the real-world applicability of the model. The institutional review boards waived the requirement for written informed consent for patients included in the retrospective cohort, while written informed consent was obtained from every patient in the prospective cohort. Women were excluded from the study if they had an intrauterine device, if the final pathology report lacked information about the endometrium, or if they had been diagnosed with any malignant tumor other than endometrial cancer. Details on the selection of ultrasound images and the corresponding endometrial pathological diagnoses were provided in the Supplementary Material. Patients diagnosed with endometrial cancer or atypical hyperplasia were categorized into the malignant group, while those with other pathological diagnoses were categorized into the non-malignant group.

All ultrasound examiners had more than 10 years of experience in gynecological ultrasound. Examinations were performed using high-performance ultrasound equipment (Hospital One: Voluson E10, Mindray R8, and HD15; Hospital Two: Voluson E10, Voluson E8, and HD15). In both hospitals, ultrasound scans were conducted with the woman in the lithotomy position and with an empty bladder, following a standardized protocol based on the IETA consensus statement to ensure consistency and reproducibility across centers [3]. The uterus was first scanned in both the sagittal plane (from cornu to cornu) and the transverse plane (from cervix to fundus). After obtaining an overview of the entire uterus, the image was magnified to focus specifically on the uterine corpus, with magnification adjusted to center on the area of interest.

Sample size determination

To determine the number of samples required, we used a publicly available web-based sample size calculator (https://wnarifin.github.io/ssc/sssnsp.html), which was based on the statistical methodology described by Buderer and developed by Arifin [19, 20]. The input parameters were set as follows: Expected sensitivity: 0.90; Expected specificity: 0.90; Prevalence of disease: 0.14; Precision (± expected): 0.05; Confidence level: 95%; and Expected dropout rate: 10%. Based on these input parameters, the estimated minimum required sample size for the testing set was 275. Assuming a training-to-testing ratio of 4:1, the total required sample size was calculated to be 1,375.

Image pre-processing

To improve image quality and enhance model performance, we applied a series of standardized pre-processing techniques to all ultrasound images. First, two gynecologists independently delineated two regions of interest (ROIs), the endometrium and the uterine corpus, on each image. Discrepancies were resolved through expert consensus, and the ROIs were cropped to remove irrelevant background information. Subsequently, three pre-processing steps were applied: (1) Denoising using the Non-local Means (NLM) algorithm to reduce speckle noise; (2) Normalization by zero-centering pixel intensities based on the global mean and standard deviation across the dataset; and (3) Super-resolution using a deep learning-based super-resolution generative adversarial network (SRGAN) model, trained on a large private ultrasound dataset, to enhance resolution by fourfold. We selected SRGAN for super-resolution processing due to its demonstrated effectiveness in enhancing medical images, particularly in preserving critical anatomical details. Previous studies have shown that SRGAN outperforms traditional methods in terms of visual fidelity and structural similarity, making it well-suited for medical imaging tasks where high-resolution detail is crucial [21]. For detailed implementation and parameters of these preprocessing techniques, please refer to the Supplementary Material. Sensitivity analyses were conducted on the training set to assess how diagnostic performance changed after image pre-processing. The images used for these analyses were rectangular regions closely surrounding the endometrium.

Radiomics feature extraction

Radiomics features were extracted using the Pyradiomics package (version 3.0.1) in Python [22]. Images were normalized with a fixed scale factor of 1000, and intensity discretization was performed using a fixed bin width of 5. Given the lack of consistent spatial resolution metadata in ultrasound images, no voxel resampling was performed, and feature extraction was conducted at the native image resolution to preserve the original texture information. Features were extracted from both the original images and selected filtered versions, including wavelet, gradient, and logarithmic transforms. For Laplacian of Gaussian (LoG) filtering, $\sigma = 1.0$ was chosen as it provides an optimal balance between smoothing and noise reduction for ultrasound images. A padding distance of 3 voxels was applied to minimize edge effects. Extracted features included shape, first-order statistics, and texture features, including Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM).

Deep learning feature extraction

Four CNNs, specifically VGG19, ResNet18, ResNet50, and Inception-v3, were selected and utilized for the construction of image-based classification models due to their superior performance in the ImageNet Large-scale Visual Recognition Challenge (ILSVRC). Each model was chosen for its unique architecture. VGG19 features a deep and homogeneous structure, effectively capturing hierarchical image features without complex modules. ResNet18 and ResNet50 utilize residual blocks to train deeper networks without the vanishing gradient problem, with ResNet18 offering computational efficiency and ResNet50 capturing more intricate patterns. Inception-v3 incorporates skip connections and multi-scale feature extraction, enhancing its ability to capture fine details and global structures in medical images. The configurations for CNNs were as follows: the loss function was binary cross-entropy, and the optimizer was Adam with a learning rate of 0.01. Models were trained with a batch size of 32 for 50 epochs, and a dropout rate of 0.5 was applied to prevent overfitting. All ultrasound images were resized to 224×224 pixels before being fed into the

models. To improve generalization performance, data augmentation was applied to the training set, including random horizontal and vertical rotation (± 20°), scaling of width and height (± 10%), horizontal and vertical flipping (probability = 0.5), and zooming within a range of 0.8 to 1.2. The average pooling layer, particularly the one nearest to the fully connected layer, plays a crucial role in extracting the most representative features from the image, and was thus utilized for deep learning variable extraction in each of the CNNs. Pre-trained weights from ImageNet were used to initialize the models, leveraging transfer learning to improve convergence speed and performance. The average pooling layer nearest to the fully connected layer was used for feature extraction in each CNN, ensuring optimal representation of image-based features.

Clinical data collection and pre-processing

The clinical and sonographic records were retrieved from the electronic medical record system and then manually checked. Following the integration of radiomics, deep learning, and clinical features, we implemented a series of procedures to reduce data dimensionality and select the most pertinent variables for the final model. These procedures encompassed imputations, normalizations, collinearity analysis, and variable selection (see Data Preprocessing Section in Supplement Material for details).

Classification task assignment

To systematically evaluate different feature extraction and classification strategies for predicting the risk of EC in postmenopausal women using ultrasound images, we developed and compared three AI models. This approach allowed us to assess the relative strengths of radiomics-based machine learning, deep learning-based classification, and their combined implementation, in order to determine whether integrating both types of features enhance diagnostic performance. Specifically: (1) The Radiomics (R) model evaluated the predictive value of handcrafted features related to tissue heterogeneity, shape, and texture extracted from ultrasound images; (2) The CNN model assessed the effectiveness of automatically learned deep features that capture complex hierarchical spatial patterns, enabling direct comparison with radiomics-based methods; (3) The DLR model integrated both deep learning and radiomic features to explore whether their combination could achieve superior classification performance over either approach alone.

Six machine learning algorithms, including logistic regression (LR), decision tree (DT), random forest (RF), support vector machines (SVM), adaptive boosting (Adaboost), and Extreme Gradient Boosting (XGBoost), were selected based on their diverse characteristics and proven effectiveness in various classification tasks. LR provides a simple and interpretable baseline for binary classification tasks, offering a foundation for comparison. DT, with its clear and interpretable structure, is particularly effective for capturing non-linear relationships in the data. To enhance robustness and prevent overfitting, we incorporated RF, an ensemble method that combines multiple decision trees, improving model generalizability. For higher-dimensional data, SVM is known for its ability to find optimal decision boundaries, making it ideal for complex classification problems. Adaboost improves model performance by focusing on difficult-to-classify samples, effectively boosting weak classifiers, which enhances accuracy. Finally, XGBoost is favored for its high efficiency, scalability, and superior predictive performance, particularly in large and complex datasets with many features. To reduce overfitting and ensure the reliability of hyperparameter tuning, hyperparameters were tuned using ten-fold cross-validation within the training dataset. A grid search strategy was employed to identify the optimal parameter combinations based on the average area under the receiver operating characteristic curve (AUROC). Subgroup analyses were also conducted based on different ROIs.

To robustly assess model performance, we employed bootstrap resampling with 1000 iterations. In each iteration, random samples from the testing set were drawn with replacement, and predictions were made using the trained model. We compared algorithms based on discrimination and calibration. Discrimination was measured using accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and AUROC, with mean values and 95% confidence intervals reported. In addition, we plotted precision-recall (PR) curves for each model and computed the area under the PR curve (AUPRC) using 1000 bootstrap iterations, reporting the mean and 95% confidence intervals. DeLong's test determined statistically significant differences between AUROC curves of different models. Calibration was visually assessed using calibration plots and quantitatively by Brier score. Decision curve analysis calculated the net benefit for potential clinical use. Descriptive statistics were performed using the "SCIPY" package in Python, with continuous variables presented as mean ± standard deviation or median (quartile [Q] 25–Q75) if not normally distributed, and categorical variables presented as number (%) with comparisons made using χ^2 tests. Continuous variables were compared using t-tests or Mann-Whitney U-tests if not normally distributed.

Results

During the study period, 2545 eligible women were identified, all having both ultrasound and endometrial pathology data available. Of these, 256 had an intrauterine

device; and 77 were diagnosed with cancers other than endometrial cancer, specifically cervical cancer ($n = 53$), primary ovarian cancers ($n = 11$), primary tubal cancers ($n = 2$), synchronous ovarian and tube cancers ($n = 2$), breast cancer ($n = 7$), colon cancer ($n = 1$), and placental site trophoblastic tumor ($n = 1$). Additionally, 236 women lacked endometrial pathology results, and the images of 115 were unavailable or unsuitable for outlining the uterus corpus or endometrium. Thus, the final study cohort comprised 1861 women. The overall workflow pattern of this study is depicted in Fig. 1 and Supplementary Figure S1. The actual sample size for the training and testing sets met the initial sample size calculation requirements, ensuring sufficient statistical power for model training and evaluation. The clinical and sonographic characteristics of postmenopausal women in the training, internal and external testing sets are summarized in Supplementary Table S1 in Supplementary Material. Notably, patients in the malignant group exhibited thicker endometrium, larger uterine volume, and a reduced prevalence of uterine fibroids when compared to the non-malignant group.

Image preprocessing

The architecture of SRGAN and examples of the enhanced images are depicted in Fig. 2. The results of sensitivity analyses are listed in Table 1. Models trained on images processed with denoising, normalization, super-resolution, or a combination of these methods performed better than those trained on raw images. However, only super-resolution and combined preprocessing achieved statistically significant improvements in model performance, with AUROC increases of 0.834 ($P < 0.05$) and 0.853 ($P < 0.01$), respectively, compared to the raw image model (AUROC 0.794). These findings highlight that appropriate image preprocessing, particularly the integration of multiple methods, can significantly enhance diagnostic accuracy in deep learning models.

Data preparation and preprocessing

To predict the risk of endometrial cancer (EC) within the endometrium-level ROI, a total of 3616 variables were retrieved, comprising 1562 radiomic features, 2048 deep learning features, and 6 clinical variables. After analyzing the variance of each variable, those with a variance less than 0.1 were removed, resulting in the elimination of 2679 variables. Following MICE imputation and data normalization, we assessed the Variance Inflation Factor (VIF) for each variable and excluded those with a VIF value greater than five, resulting in 793 variables being removed. Finally, 36 variables (18 radiomic, 16 deep learning, and 2 clinical) were selected for the DLR model construction using the least absolute shrinkage and selection operator (LASSO) regression, based on the remaining 144 variables with a VIF less than five (Fig. 3a and b).

When predicting the risk of EC based on a uterine-corpus-level ROI, 3616 variables were retrieved (note that the number of radiomic and deep learning features extracted from the uterine-corpus level is identical to that from the endometrium-level ROI but differs significantly in contextual content). After preprocessing, 12 variables were selected for model construction (Fig. 3c and d).

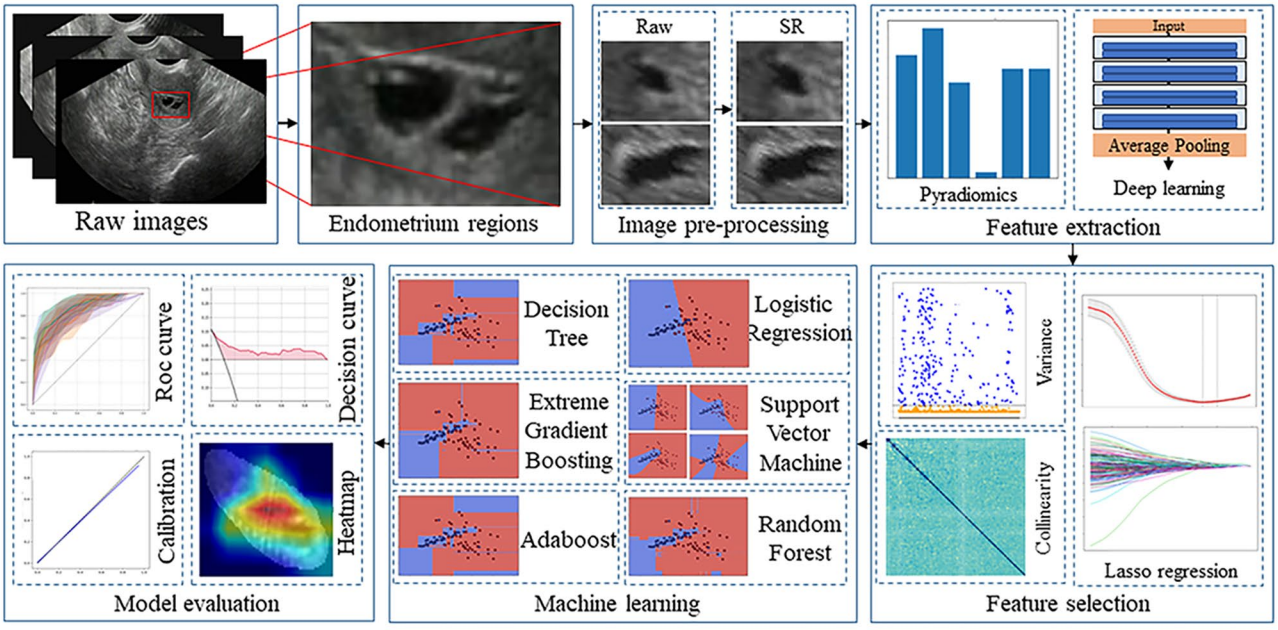

Fig. 1 The flow chart of the present study

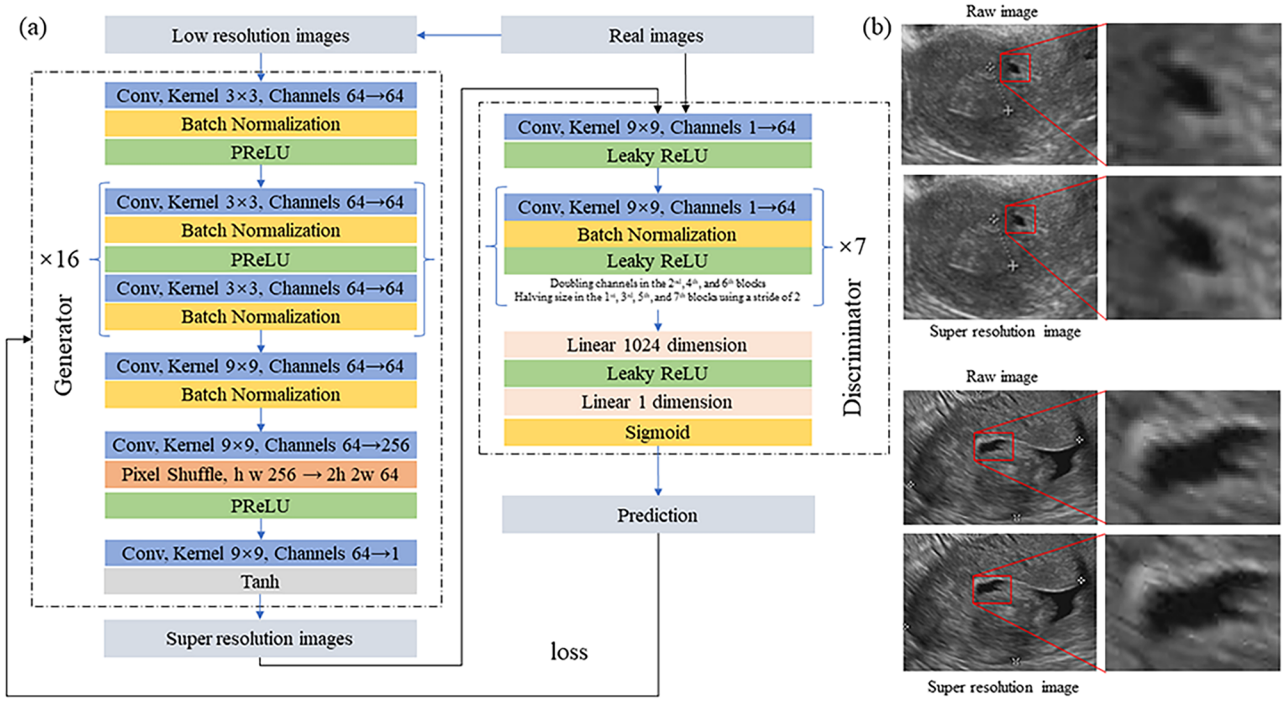

Fig. 2 Diagram of generative adversarial networks used to generate super-resolution ultrasound images from normal-resolution ultrasound images. (**a**) The architecture of super resolution generative adversarial networks; (**b**) Two examples for ultrasound images before and after super resolution

Table 1 Comparison of the diagnostic performances among model (ResNet-50) under various image pre-processing

Image preprocessing	Training	Test	Accuracy	Sensitivity	Specificity	PPV	NPV	AUROC	*P*-value*
Denoising	Raw	Raw	0.771 (0.747,0.795)	0.680 (0.645,0.716)	0.787 (0.757,0.817)	0.386 (0.356,0.417)	0.931 (0.924,0.937)	0.794 (0.770,0.817)	Ref
	Denoised	Denoised	0.786 (0.761,0.810)	0.688 (0.648,0.728)	0.804 (0.771,0.837)	0.411 (0.369,0.454)	0.934 (0.928,0.940)	0.809 (0.790,0.827)	> 0.05
Normalization	Raw	Raw	0.771 (0.747,0.795)	0.680 (0.645,0.716)	0.787 (0.757,0.817)	0.386 (0.356,0.417)	0.931 (0.924,0.937)	0.794 (0.770,0.817)	Ref
	Normalized	Normalized	0.794 (0.758,0.829)	0.704 (0.651,0.756)	0.810 (0.763,0.857)	0.418 (0.366,0.471)	0.937 (0.928,0.947)	0.820 (0.796,0.844)	> 0.05
SR	Raw	Raw	0.771 (0.747,0.795)	0.680 (0.645,0.716)	0.787 (0.757,0.817)	0.386 (0.356,0.417)	0.931 (0.924,0.937)	0.794 (0.770,0.817)	Ref
	SR	SR	0.789 (0.750,0.827)	0.717 (0.675,0.760)	0.802 (0.750,0.853)	0.422 (0.346,0.498)	0.940 (0.934,0.946)	0.834 (0.817,0.850)	< 0.05
Combined	Raw	Raw	0.771 (0.747,0.795)	0.680 (0.645,0.716)	0.787 (0.757,0.817)	0.386 (0.356,0.417)	0.931 (0.924,0.937)	0.794 (0.770,0.817)	Ref
	Combined	Combined	0.811 (0.780,0.842)	0.737 (0.699,0.775)	0.824 (0.783,0.866)	0.447 (0.396,0.498)	0.945 (0.938,0.952)	0.853 (0.839,0.868)	< 0.01

* P-values are calculated using the DeLong test to compare the AUC values between the respective image preprocessing method (Denoising, Normalization, Super Resolution, Combined) and the raw data. A P-value less than 0.05 indicates a significant difference in AUC between the processed and raw data for each preprocessing method

NPV: negative predictive value; PPV: positive predictive value; SR: super resolution

The performance of models among postmenopausal patients

The DLR models developed to predict the risk of EC using endometrium-level ROIs were evaluated and compared (Table 2). Hyperparameter settings are summarized in Supplementary Table S2. Among all models, the SVM showed the highest performance in the internal testing dataset, with a sensitivity of 0.847 (95% CI: 0.692–0.944), specificity of 0.810 (95% CI: 0.717–0.910), and AUROC of 0.893 (95% CI: 0.847–0.932) (Fig. 4a). Compared to SVM, the AUROC of all other models was significantly lower (all $P < 0.05$, DeLong test), confirming its superior discriminative ability. Similar performance was observed in the external testing dataset, where the SVM achieved a sensitivity of 0.792 (95% CI: 0.622–0.955), specificity of 0.829 (95% CI: 0.644–0.936),

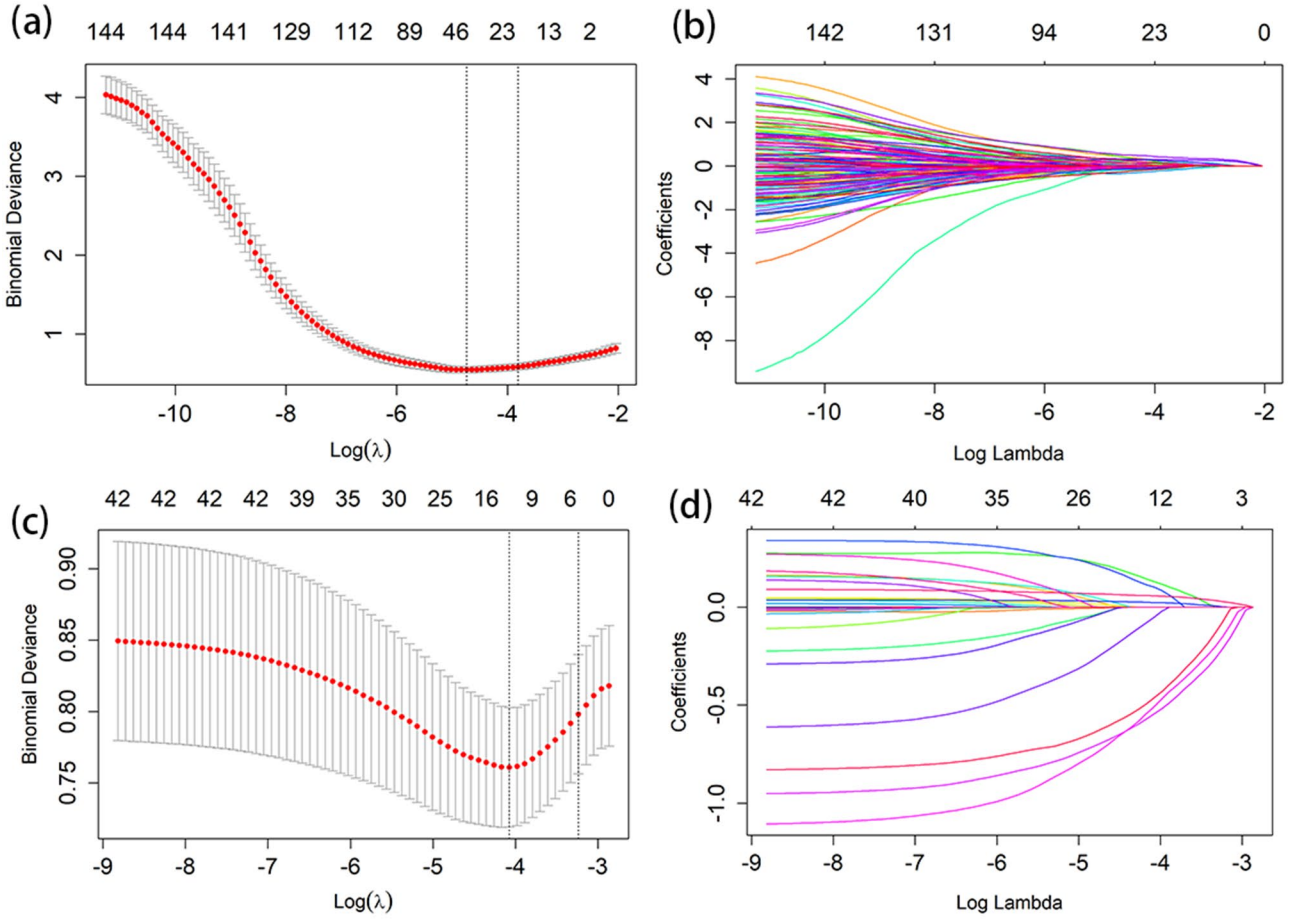

Fig. 3 Variable selection using the LASSO binary logistic regression model. (**a**) Identification of the optimal penalization coefficient lambda (λ) in the LASSO model with 10-fold cross-validation when predicting risk of EC using endometrium-level ROIs. (**b**) Coefficient profiles of 144 variables predicting EC using endometrium-level ROIs were exhibited by LASSO logistic regression, based on the log(λ) sequence. (**c**) Identification of the optimal penalization coefficient lambda (λ) in the LASSO model with a 10-fold ross-validation when predicting risk of EC using uterine-corpus-level ROIs. (**d**) Coefficient profiles of 42 variables predicting EC using uterine-corpus-level ROIs were exhibited by LASSO logistic regression, based on the log(λ) sequence. LASSO, least absolute shrinkage and selection operator; EC, endometrial cancer

and AUROC of 0.871 (95% CI: 0.804–0.930) (Fig. 4b). Pairwise comparisons of AUROCs showed that SVM significantly outperformed most models ($P<0.05$), except XGBoost ($P>0.05$) and LR ($P>0.05$) in the external dataset. The SVM model demonstrated the highest AUPRC in both internal (AUPRC = 0.660, 95% CI: 0.532–0.770) and external (AUPRC = 0.649, 95% CI: 0.501–0.775) testing datasets (Table 1; Fig. 4c and d). Decision curve analysis revealed that our model provided a higher net benefit compared to both intervening for all patients and not intervening for any patients (Supplementary Figure S2a and S2b). The calibration curves of the SVM model in both testing datasets are shown in Supplementary Figure S2c and S2d. The results of the R models were listed in Supplementary Table S3, with ROC curves, calibration curves, and decision curves depicted in Supplementary Figure S3 and Supplementary Figure S4. The results of CNN models were listed in Supplementary Tables S4 and Supplementary Figure S5.

The performance of the DLR model was superior to that of the R and CNN models across both the internal and external testing datasets (Table 3). In the internal testing dataset, the DLR model achieved the highest AUROC of 0.893 (95% CI: 0.847–0.932), which was significantly higher than that of the R model (AUROC = 0.778, 95% CI: 0.711–0.839; $P<0.001$) and the CNN model (AUROC = 0.828, 95% CI: 0.700–0.856; $P<0.01$). Similarly, in the external testing dataset, the DLR model (AUROC = 0.871, 95% CI: 0.804–0.930) outperformed both the R model (AUROC = 0.785, 95% CI: 0.710–0.857; $P<0.001$) and the CNN model (AUROC = 0.798, 95% CI: 0.620–0.976; $P<0.001$), with statistically significant differences based on DeLong's test.

When using the uterine-corpus-level ROI, the logistic regression model exhibited superior discrimination in the external testing dataset (Supplementary Table S5). Overall, the performance of DLR models based on the

Table 2 Discrimination metrics of DLR models to predict the risk of endometrial cancer among postmenopausal women with endometrium-level regions of interests

Datasets	Models	Performance matrix						P-value*
		Accuracy	Sensitivity	Specificity	PPV	NPV	AUROC	
The internal testing dataset	LR	0.800 (0.563,0.920)	0.743 (0.553,1)	0.809 (0.505,0.960)	0.451 (0.2,0.707)	0.959 (0.925,1)	0.850 (0.798,0.895)	< 0.05
	RF	0.832 (0.673,0.908)	0.637 (0.453,0.844)	0.860 (0.654,0.956)	0.440 (0.235,0.667)	0.942 (0.916,0.970)	0.797 (0.728,0.857)	< 0.001
	SVM	0.815 (0.741,0.892)	0.847 (0.692,0.944)	0.810 (0.717,0.910)	0.404 (0.295,0.581)	0.973 (0.951,0.990)	0.893 (0.847,0.932)	Ref
	XGBoost	0.781 (0.565,0.858)	0.739 (0.591,0.930)	0.787 (0.515,0.884)	0.352 (0.217,0.478)	0.954 (0.929,0.981)	0.817 (0.756,0.872)	< 0.05
	DT	0.806 (0.730,0.847)	0.573 (0.435,0.709)	0.840 (0.732,0.880)	0.347 (0.247,0.447)	0.931 (0.902,0.956)	0.700 (0.620,0.774)	< 0.001
	Adaboost	0.802 (0.590,0.902)	0.666 (0.475,0.891)	0.822 (0.558,0.943)	0.400 (0.207,0.619)	0.945 (0.911,0.975)	0.787 (0.720,0.853)	< 0.001
The external testing dataset	LR	0.788 (0.635,0.942)	0.795 (0.554,0.972)	0.787 (0.599,0.982)	0.360 (0.190,0.793)	0.970 (0.942,0.995)	0.849 (0.772,0.910)	> 0.05
	RF	0.757 (0.484,0.920)	0.753 (0.511,1)	0.758 (0.422,0.960)	0.342 (0.146,0.652)	0.965 (0.930,1)	0.828 (0.759,0.889)	< 0.01
	SVM	0.825 (0.670,0.915)	0.792 (0.622,0.955)	0.829 (0.644,0.936)	0.400 (0.217,0.603)	0.971 (0.945 0.993)	0.871 (0.804,0.930)	Ref
	XGBoost	0.788 (0.602,0.909)	0.808 (0.628, 1)	0.786 (0.558,0.938)	0.343 (0.194,0.576)	0.972 (0.946, 1)	0.855 (0.790,0.913)	> 0.05
	DT	0.734 (0.63, 0.841)	0.756 (0.552,0.893)	0.732 (0.659,0.862)	0.264 (0.178,0.392)	0.961 (0.932,0.984)	0.770 (0.696,0.836)	< 0.001
	Adaboost	0.790 (0.626,0.923)	0.750 (0.510,0.939)	0.795 (0.604,0.960)	0.355 (0.181,0.658)	0.964 (0.935,0.990)	0.831 (0.750,0.894)	< 0.05

* P-values are calculated using the DeLong test to compare the AUC values between SVM and other models within each dataset. A P-value less than 0.05 indicates a significant difference in AUC between SVM and the other model

AUROC: area under the receiver operating characteristic curve; Adaboost: adaptive boosting; DT: decision tree; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting

endometrium-level ROI was superior to that based on the uterine-level ROI.

Variable importance

The top three contributing variables included a DL feature (ResNet_1179), a radiomic feature (wavelet_HLL_firstorder_Mean), and a clinical feature (age) (Fig. 5a). The SHapley Additive exPlanations (SHAP) values for three individual patients are shown in Fig. 5b, demonstrating how each feature contributed to the prediction outcome for that sample. Positive SHAP values indicate an increased likelihood of EC, while negative values suggest a reduced likelihood. Figure 5c displays the SHAP force plot for the internal test dataset. The horizontal axis shows samples ordered by similarity in feature contribution patterns, and the vertical axis reflects the cumulative SHAP values, indicating the model output (f(x)).

Discussion

Our findings in this study suggest that ultrasound-based model, incorporating deep learning and radiomics features, can be reliable in predicting the risk of endometrial cancer. Such models could be used for the early screening of endometrial cancer solely on ultrasound images.

The clinical implication of our study is the development of a high-performing DLR model based solely on ultrasound images for the early diagnosis of EC. Traditional ultrasound screening primarily relies on ET, which has poor specificity (51.5%) at a 5 mm threshold, often leading to unnecessary follow-ups [23, 24]. Several advanced ultrasound techniques, including endometrial morphological assessment and Doppler imaging, have been explored to improve diagnostic accuracy. For example, the interrupted endo-myometrial junction showed an AUROC of 0.70, with sensitivity of 62% and specificity of 78%, while Doppler imaging demonstrated an AUROC of 0.745, with sensitivity of 72.4% and specificity of 74.4% [25]. Despite these advancements, these methods remain highly operator-dependent, as evidenced by significant variability in inter- and intra-observer agreement across different ultrasound parameters, limiting their feasibility for widespread clinical adoption [26]. Several AI-based approaches have been explored for the diagnosis of EC using ultrasound images [27]. The Risk of Endometrial Cancer (REC) score, a model combining ultrasound-derived features and clinical parameters through logistic regression, achieved an AUROC of 0.75 (95% CI: 0.70–0.79), with a sensitivity of 79% and specificity of

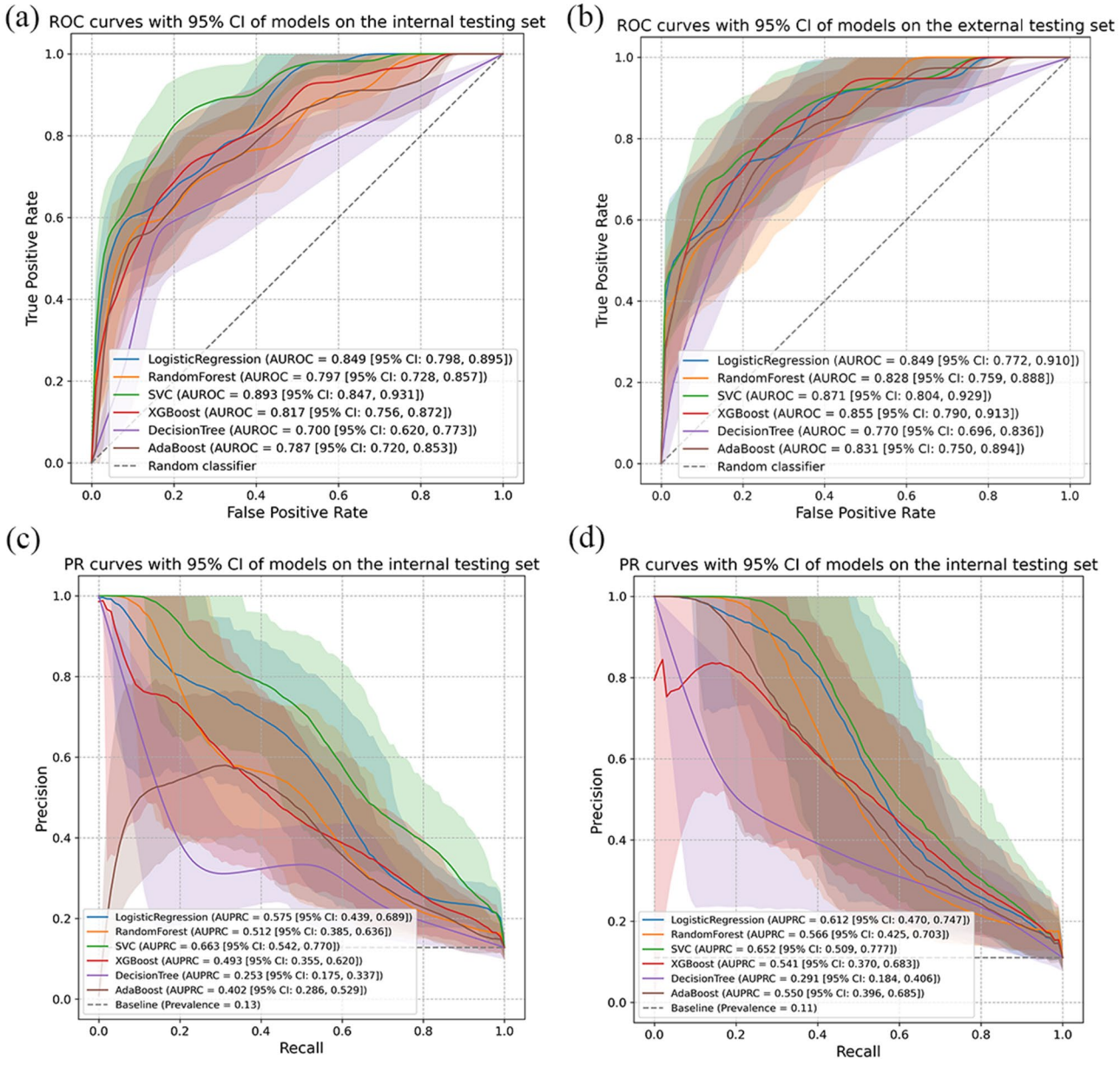

Fig. 4 Model diagnostics for DLR models using endometrium-level ROIs on the internal and external testing sets. (**a**) ROC curves of each machine learning model on the internal testing sets; (**b**) ROC curves of each machine learning model on the external testing sets; (**c**) PR curves of each machine learning model on the internal testing sets; (**d**) PR curves of each machine learning model on the external testing sets

61% [28]. However, this model is limited by its reliance on predefined metrics and its inability to fully capture the complex spatial and textural patterns inherent in ultrasound images. More recently, a radiomics-based model reported an AUROC of 0.90 in the validation set and 0.88 in the test set for EC diagnosis, further demonstrating the potential of AI-based approaches in this field [29]. Although our DLR model exhibited comparable diagnostic performance (AUROC = 0.893 in the internal testing set and 0.871 in the external testing set), our study differs in several important aspects, particularly in terms of study population and methodology. First, while the referenced study focused on patients with postmenopausal bleeding, a population with a higher pretest probability of malignancy, our study included a broader cohort of postmenopausal women, enhancing the generalizability of our findings for early detection scenarios. Second, we employed SR techniques to enhance image quality, which was not implemented in the previous study. Unlike CT and MRI, which are not typically used for EC screening, ultrasound is the primary imaging modality in this context. Therefore, our comparisons have focused on ultrasound-based AI approaches, which are most relevant to the intended clinical application.

Table 3 Comparison of top-performing algorithms in DLR, R, and CNN models among postmenopausal women with uterine-level regions of interests

Datasets	Models[&]	Performance matrix						P-value*
		Accuracy	Sensitivity	Specificity	PPV	NPV	AUROC*	
The internal testing dataset	DLR	0.815 (0.741,0.892)	0.847 (0.692,0.944)	0.810 (0.717,0.910)	0.404 (0.295,0.581)	0.973 (0.951,0.990)	0.893 (0.847,0.932)	Ref
	R	0.767 (0.636,0.888)	0.693 (0.479,0.873)	0.780 (0.605,0.943)	0.383 (0.233,0.625)	0.938 (0.904,0.970)	0.778 (0.711,0.839)	< 0.001
	CNN	0.777 (0.657,0.897)	0.737 (0.591,0.884)	0.784 (0.626,0.943)	0.404 (0.251,0.557)	0.943 (0.918,0.967)	0.828 (0.700,0.856)	< 0.01
The external testing dataset	DLR	0.825 (0.670,0.915)	0.792 (0.622,0.955)	0.829 (0.644,0.936)	0.400 (0.217,0.603)	0.971 (0.945 0.993)	0.871 (0.804,0.930)	Ref
	R	0.782 (0.662,0.879)	0.716 (0.531,0.857)	0.792 (0.642,0.936)	0.364 (0.245,0.538)	0.947 (0.917,0.971)	0.785 (0.710,0.857)	< 0.001
	CNN	0.765 (0.639,0.891)	0.698 (0.502,0.894)	0.777 (0.614,0.941)	0.384 (0.225,0.543)	0.934 (0.899,0.970)	0.798 (0.620,0.976)	< 0.001

* The P-values represent the statistical significance of the performance differences between each model and DLR, calculated using the DeLong test. A P-value less than 0.05 indicates a significant difference in AUC between DLR and the other model

[&] The models shown in the table represent the top-performing algorithms for each model type (DLR, R, and CNN) based on internal and external testing datasets

AUROC: area under the receiver operating characteristic curve; CNN: convolutional neural network; DLR: deep learning radiomics model; R: radiomics model

Another implication is the importance of carefully selecting the appropriate ROI to optimize model performance. Theoretically, the uterine ROI should provide more comprehensive information, as it encompasses both the endometrium and the endometrial–myometrial junction [23]. However, our results showed that the model using the endometrial ROI outperformed the one based on the uterine ROI. This discrepancy suggests that additional anatomical structures, such as uterine fibroids or adenomyosis, may introduce noise and confounding factors, disrupting feature extraction and hindering the model's ability to accurately predict endometrial cancer. Therefore, restricting the ROI to the endometrium may reduce this noise and improve diagnostic accuracy, aligning with other studies focusing on EC detection using ultrasound images [29].

In medical imaging, especially ultrasound, acquiring high-resolution images can be costly and time-consuming [30]. As a result, SRGAN has emerged as a key tool for enhancing image resolution, as demonstrated in various studies utilizing publicly available medical image datasets [31]. However, significant differences exist between natural and biomedical images in SR development. Biomedical images typically exhibit lower contrast, higher noise levels, and more complex structures [32]. To address these challenges, we trained a customized SRGAN model on a private dataset of 34,117 uterine ultrasound images from the Voluson-E10 machine, ensuring no overlap with our study data. Sensitivity analysis showed that this domain-specific model substantially improved downstream performance, emphasizing the need for tailored SR approaches in clinical imaging tasks [33].

A key strength of our study is the inclusion of both a prospective testing set from the same hospital and an external testing set from a different hospital in the same city. While this design enhances the generalizability of our findings, it inevitably introduces potential sources of bias and data heterogeneity. Specifically, differences in ultrasound equipment, operator experience, image quality, and annotation consistency may exist between retrospective and prospective data collection, and between different institutions. Additionally, prospective data are generally more standardized, while retrospective data may suffer from selection bias and missing information. To mitigate these biases, we implemented a unified imaging protocol and applied a consistent preprocessing pipeline to all images prior to analysis. Importantly, the model achieved stable and robust performance across both internal and external testing sets, demonstrating good generalizability despite the inherent variability. Nonetheless, we acknowledge that such heterogeneity reflects real-world clinical settings, and future large-scale, multicenter, prospective studies are warranted to further evaluate the model's impact on clinical decision-making and patient outcomes.

Although DL approaches are often considered computationally intensive, our study demonstrated that accurate performance could be achieved with relatively modest computational resources. All analyses were performed on a single consumer-grade GPU (RTX 3090), and inference times were fast (approximately 0.2 s per image), supporting the potential for real-world clinical deployment of our models without significant computational burden. Furthermore, the model's relatively small file size and low computational requirements make it highly suitable for integration into clinical workflows. It can be deployed via a mobile app developed using Android Studio or a web application through Docker containers. This flexibility ensures that the model can be run efficiently on standard devices, such as smartphones or desktops, without

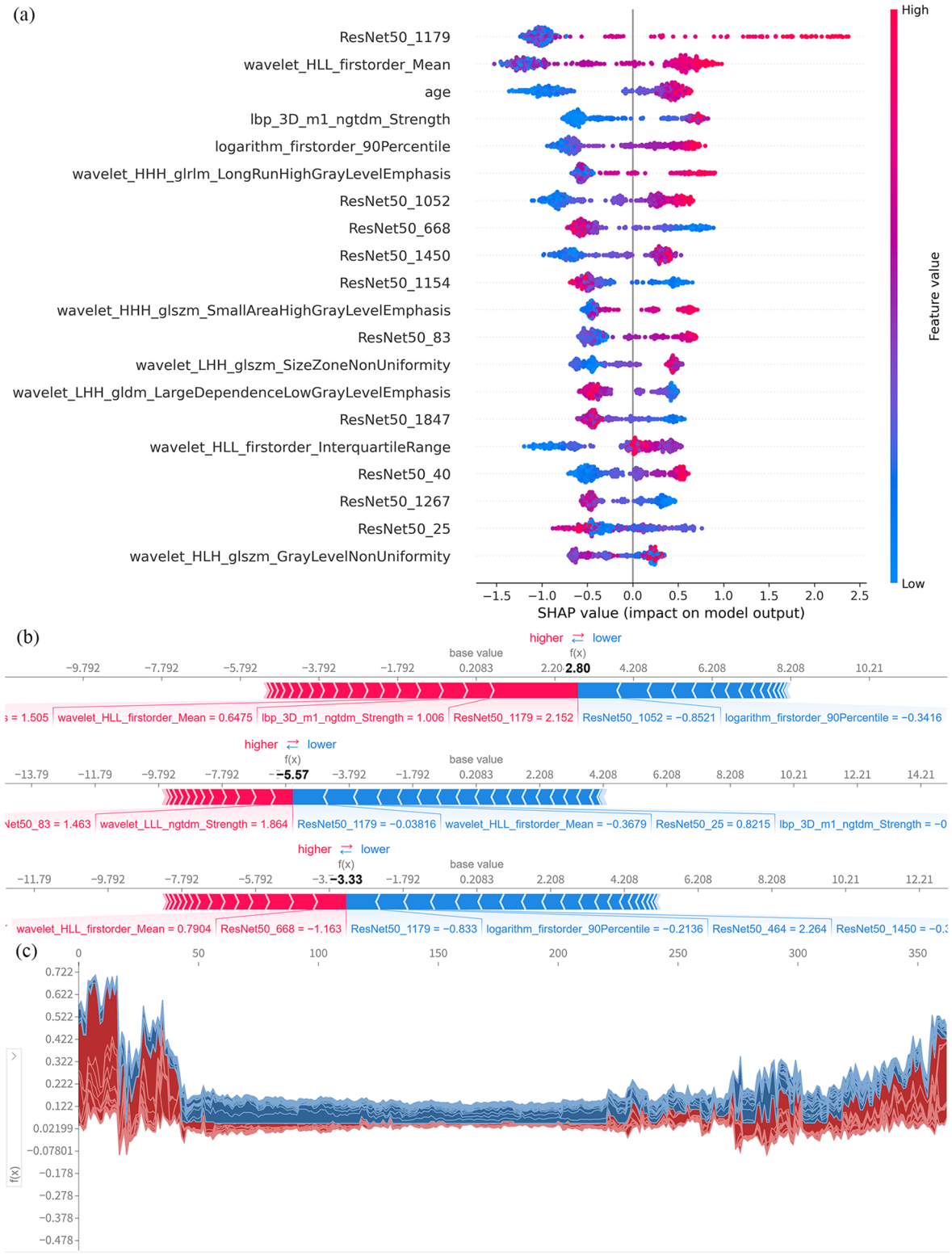

Fig. 5 Variable importance and SHAP analysis for predicting endometrial cancer. (**a**) variable importance; (**b**) SHAP values for three specific patients; (**c**) SHAP force plot for the entire internal testing dataset

significant resource consumption. The model's fast inference time supports its potential for real-time clinical deployment, providing clinicians with timely diagnostic assistance without introducing substantial computational burden or privacy concerns. Given these features, our model could be easily integrated into clinical settings, enhancing diagnostic workflows and supporting decision-making in a wide range of healthcare environments.

However, our study has certain limitations. First, key clinical variables such as obesity and nulliparity were not available in our dataset. These are established risk factors for EC, primarily due to their association with increased lifetime exposure to unopposed estrogen, a major contributor to endometrial carcinogenesis [34]. Additionally, hormone therapy usage, another important clinical variable, was not included in our dataset, which could have influenced the model's ability to fully capture individualized risk profiles. The absence of these variables, including hormone therapy usage, may have limited the model's ability to account for a more complete set of risk factors, potentially affecting prediction accuracy. Second, due to the retrospective design of the study, other clinically relevant factors such as diabetes, polycystic ovary syndrome (PCOS), and hormone therapy use were either not recorded or present in too few cases to allow for meaningful subgroup analyses. This further limits the generalizability of the model, especially in subgroups where these clinical variables play a key role, such as in patients with obesity or those undergoing hormone therapy, and affects our ability to evaluate its performance across diverse clinical conditions [35]. Third, our analysis was based solely on grayscale transvaginal ultrasound images. While grayscale imaging is routinely used in clinical practice for endometrial evaluation, it does not capture blood flow information that can be obtained through Doppler imaging. We acknowledge the potential added diagnostic value of Doppler imaging in assessing vascular patterns and are currently planning a prospective study that will integrate Doppler imaging and additional clinical and laboratory parameters to further enhance model performance and clinical applicability.

In summary, our ultrasound-based machine learning model, which seamlessly integrates DL and radiomics features, effectively supports the early screening of EC. Specifically, in the two testing datasets, our model achieved an AUC of 0.893 (95%CI: 0.847,0.932) and 0.871 (95%CI: 0.804,0.930), respectively. Compared to existing methods, these results demonstrate the model's ability to accurately identify EC risk at an early stage using ultrasound images.

Abbreviations

Adaboost	Adaptive boosting
AUROC	Area under the receiver operating characteristic curve
CNN	Convolutional neural networks
DT	Decision tree
DL	Deep learning
DLR	Deep learning and radiomics
EC	Endometrial cancer
ET	Endometrial thickness
ML	Machine learning
PCOS	polycystic ovary syndrome
ROI	Region of interest
RF	Random forests
REC	Risk of endometrial cancer score
SHAP	SHapley additive exPlanations
SMOTE	Synthetic minority oversampling technique
SR	Super resolution
SRGAN	Super resolution generative adversarial network
SVM	Support vector machines
XGBoost	Extreme gradient boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-025-01705-1.

Supplementary Material 1

Acknowledgements

N/A.

Author contributions

Yi-xin Li and Yu Lu contributed equally to this research. Yi-xin Li, Yu Lu, and Min Ren conceived the study design; Yi-xin Li, Yu Lu, Zhe-ming Song, and Yu-ting Shen contributed to data collection, analysis, and interpretation; Yi-xin Li, Yu Lu, and Zhe-ming Song were involved in drafting the manuscript; Wen Lu and Min Ren supervised the research. All authors read and approved the final manuscript.

Funding

The study is supported by the Natural Science Foundation of Shanghai Municipality (Grant number: 22ZR1449400) and Shanghai Pudong New Area Health Commission (Grant number: PKJ2022-Y15).

Data availability

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Review Committee of Shanghai First Maternity and Infant Hospital, School of Medicine, Tongji University, Shanghai (approval number KS24284) and Chongming Hospital Affiliated to Shanghai University of Medicine and Health Sciences (approval number CMEC-2024-KT-03). This study only involved the analysis of imaging data, and no direct experiments on humans or human tissue samples were conducted. All procedures were performed in accordance with the Declaration of Helsinki and relevant institutional and national ethical guidelines and regulations. The institutional review boards waived the requirement for written informed consent for patients included in the retrospective cohort, while written informed consent was obtained from every patient in the prospective cohort.

Consent for publication

N/A.

Competing interests

The authors declare no competing interests.

Author details

[1]Shanghai Key Laboratory of Maternal Fetal Medicine, Shanghai Institute of Maternal-Fetal Medicine and Gynecologic Oncology, Shanghai First Maternity and Infant Hospital, School of Medicine, Tongji University, Shanghai, China

[2]Department of Gynecology, Chongming Hospital Affiliated to Shanghai University of Medicine and Health Sciences, Shanghai, China
[3]Department of Ultrasound, Shanghai First Maternity and Infant Hospital, School of Medicine, Tongji University, Shanghai, China

Received: 26 January 2025 / Accepted: 5 May 2025
Published online: 01 July 2025

References

1. Sung H, et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Makker V, et al. Endometrial cancer. Nat Rev Dis Primers. 2021;7(1):88.
3. Leone FP, et al. Terms, definitions and measurements to describe the sonographic features of the endometrium and intrauterine lesions: a consensus opinion from the international endometrial tumor analysis (IETA) group. Ultrasound Obstet Gynecol. 2010;35(1):103–12.
4. Dueholm M, et al. Diagnostic methods for fast-track identification of endometrial cancer in women with postmenopausal bleeding and endometrial thickness greater than 5 mm. Menopause. 2015;22(6):616–26.
5. Wong M, et al. Ultrasound diagnosis of endometrial cancer by subjective pattern recognition in women with postmenopausal bleeding: prospective inter-rater agreement and reliability study. Ultrasound Obstet Gynecol. 2021;57(3):471–7.
6. Ince S, et al. Deep learning for cerebral vascular occlusion segmentation: A novel ConvNeXtV2 and GRN-integrated U-Net framework for diffusion-weighted imaging. Neuroscience. 2025;574:42–53.
7. İnce S, Kunduracioglu I, Bayram B, Pacal I. U-Net-Based models for precise brain stroke segmentation. Chaos Theory Appl. 2025;7(1):50–60.
8. Ozdemir B, Aslan E, Pacal I. Attention enhanced InceptionNeXt-Based hybrid deep learning model for lung Cancer detection. IEEE Access 13.
9. Pacal I et al. A novel CNN-ViT-based deep learning model for early skin cancer diagnosis. Biomed Signal Process Control, 2025:104(000).
10. Chen M, et al. Development and validation of convolutional neural network-based model to predict the risk of Sentinel or non-sentinel lymph node metastasis in patients with breast cancer: a machine learning study. EClinicalMedicine. 2023;63:102176.
11. Cai G, et al. Artificial intelligence-based models enabling accurate diagnosis of ovarian cancer using laboratory tests in China: a multicentre, retrospective cohort study. Lancet Digit Health. 2024;6(3):e176–86.
12. Moro F, et al. Developing and validating ultrasound-based radiomics models for predicting high-risk endometrial cancer. Ultrasound Obstet Gynecol. 2022;60(2):256–68.
13. Li YX, et al. Convolutional neural networks for classifying cervical Cancer types using histological images. J Digit Imaging. 2023;36(2):441–9.
14. Li X, Yang L, Jiao X. Comparison of traditional radiomics, deep learning radiomics and fusion methods for axillary lymph node metastasis prediction in breast Cancer. Acad Radiol. 2023;30(7):1281–7.
15. Zhang Z, et al. Value of radiomics and deep learning feature fusion models based on dce-mri in distinguishing sinonasal squamous cell carcinoma from lymphoma. Front Oncol. 2024;14:1489973.
16. Dicle O. Artificial intelligence in diagnostic ultrasonography. Diagn Interv Radiol. 2023;29(1):40–5.
17. Liu T et al. Super-resolution reconstruction of ultrasound image using a modified diffusion model. Phys Med Biol, 2024:69(12).
18. Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
19. Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med. 1996;3(9):895–900.
20. Wan Nor Arifin NM. Sample size determination for machine learning in medical research. 2025.
21. Sood R et al. An Application of Generative Adversarial Networks for Super Resolution Medical Imaging. 2019.
22. van Timmeren JE, et al. Radiomics in medical imaging—how-to guide and critical reflection. Insights into Imaging. 2020;11(1):91.
23. Vitale SG, et al. Risk of endometrial cancer in asymptomatic postmenopausal women in relation to ultrasonographic endometrial thickness: systematic review and diagnostic test accuracy meta-analysis. Am J Obstet Gynecol. 2023;228(1):22–e352.
24. Long B, et al. Ultrasound detection of endometrial cancer in women with postmenopausal bleeding: systematic review and meta-analysis. Gynecol Oncol. 2020;157(3):624–33.
25. Liu MJ, et al. Application of transvaginal three-dimensional power doppler ultrasound in benign and malignant endometrial diseases. Med (Baltim). 2019;98(46):e17965.
26. Dueholm M, et al. An ultrasound algorithm for identification of endometrial cancer. Ultrasound Obstet Gynecol. 2014;43(5):557–68.
27. Moro F, et al. Role of artificial intelligence applied to ultrasound in gynecology oncology: A systematic review. Int J Cancer. 2024;155(10):1832–45.
28. Stachowicz N et al Risk assessment of endometrial hyperplasia or endometrial Cancer with simplified Ultrasound-Based scoring systems. Diagnostics (Basel), 2021:11(3).
29. Capasso I, et al. Artificial intelligence model for enhancing the accuracy of transvaginal ultrasound in detecting endometrial cancer and endometrial atypical hyperplasia. Int J Gynecol Cancer. 2024;34(10):1547–55.
30. Varshitha S et al. Enhancing Medical Imaging Resolution: Exploring SRGAN for High-Quality Medical Image Reconstruction. in. 2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE).
31. Ahmad W, et al. A new generative adversarial network for medical images super resolution. Sci Rep. 2022;12(1):9533.
32. Shin M, et al. Super-resolution techniques for biomedical applications and challenges. Biomed Eng Lett. 2024;14(3):465–96.
33. Gu Y, et al. MedSRGAN: medical images super-resolution using generative adversarial networks. Multimedia Tools Appl. 2020;79(29):21815–40.
34. Katagiri R, et al. Reproductive factors and endometrial Cancer risk among women. JAMA Netw Open. 2023;6(9):e2332296.
35. Drab A, et al. Evaluation of the impact of diabetes mellitus on endometrial cancer risk: an updated meta-analysis of case-control studies. Arch Med Sci. 2024;20(6):2056–62.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.