Am J Cancer Res 2024;14(12):5769-5783
www.ajcr.us /ISSN:2156-6976/ajcr0160797

Original Article

XGBoost-based nomogram for predicting lymph node metastasis in endometrial carcinoma

Xiaoting Lin[1], Fumin Gao[2], Haijiao Lin[3], Wang Yao[1], Yuxia Wang[1]

[1]Department of Reproductive Medicine, The First Affiliated Hospital, Jinan University, Guangzhou 510000, Guangdong, China; [2]Guangzhou Key Laboratory of Metabolic Diseases and Reproductive Health, Guangdong-Hong Kong Metabolism and Reproduction Joint Laboratory, Reproductive Medicine Center, Guangdong Second Provincial General Hospital, Guangzhou 510000, Guangdong, China; [3]Department of Pediatrics, The Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou 510000, Guangdong, China

Received September 27, 2024; Accepted December 4, 2024; Epub December 15, 2024; Published December 30, 2024

Abstract: This study aims to construct and optimize risk prediction models for lymph node metastasis (LNM) in endometrial carcinoma (EC) patients, thus improving the identification of patients at high risk of LNM and further providing accurate support for clinical decision-making. This retrospective analysis included 541 cases of EC treated at The First Affiliated Hospital, Jinan University between January 2017 and January 2022. Various clinical and pathological variables were incorporated, including age, body mass index (BMI), pathological grading, myometrial invasion, lymphovascular space invasion (LVSI), estrogen receptor (ER) and progesterone receptor (PR) levels, and tumor size. Multivariate Logistic regression analysis was used to identify independent risk factors for LNM. Subsequently, the Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting (XGBoost), RandomForest, and Support Vector Machine (SVM), all machine-learning algorithms, were adopted to select features and build models. The XGBoost model gave the best performance among all models, with areas under the curve (AUCs) of 0.876 and 0.832 for training and validation sets, respectively, suggesting its high discriminatory ability and prediction accuracy. Moreover, the calibration curve analysis further verified the consistency of the model-predicted values with the actual results, indicating the model's good applicability at various risk levels. According to the decision curve analysis, the XGBoost model showed high net benefits within most risk-threshold ranges, indicating its substantial practical value in clinical applications. Conclusively, this study successfully builds machine-learning models based on multiple clinical and pathological features, which can effectively predict the LNM risk in EC patients. The model is expected to provide important references for clinicians in surgical decision-making and the formulation of individualized treatment plans, thereby enhancing patient outcomes.

Keywords: XGBoost regression model, endometrial carcinoma, lymph node metastasis, prediction models

Introduction

Endometrial carcinoma (EC) is one of the most commonly malignancies of the female reproductive system worldwide, ranking second in prevalence among such malignancies in China [1]. According to global cancer statistics, around 420,000 new EC cases were reported globally in 2020, accounting for 4.4% of all newly diagnosed cancer cases among women [2]. The annual EC incidence in China is approximately 16 per 100,000 women, with a notable rise in recent years, especially among younger women [3]. Research indicates that a woman's lifetime risk of developing EC is about 3%, a figure that continues to increase globally [4]. EC has a high metastasis rate, with lymphatic metastasis being one of the most significant routes. Lymph node (LN) metastasis (LNM) not only influences the pathological staging and treatment decisions but also plays a critical role in prognosis and survival rates [5]. Consequently, analyzing the influencing factors of LNM is of great significance for evaluating patients' prognoses and determining the necessity of LN dissection.

Currently, total abdominal hysterectomy plus bilateral adnexectomy and evaluation of LNs (pelvic and para-aortic LNs) is the standard sur-

https://doi.org/10.62347/JVRG8195

gical modality for EC patients [6]. Precise LN assessment and dissection are crucial for determining the extent of the lesion, accurate staging, prognostic evaluation, and guiding subsequent treatment. However, the clinical value of lymphadenectomy remains somewhat controversial [7]. To a certain extent, lymphadenectomy does not reduce survival but may heighten the risk of intraoperative and postoperative complications, such as vascular injury, lymphocele, and lower-limb lymphedema, in EC patients [8]. Since the risks and benefits of lymphadenectomy vary among patients, effectively identifying high-risk factors for LNM to help clinicians select the appropriate scope of surgery is a key focus of research [9]. The prognosis of EC is affected by multiple factors, including pathological type, age, and molecular characteristics. Traditionally, EC is classified as either estrogen- or nonestrogen-dependent [10].

Despite the demonstrated potential of molecular typing for prognostic evaluation, its widespread adoption in China has been limited, primarily due to the nascent stage of its clinical application and challenges related to the standardization of detection methods and economic factors [11]. Consequently, timely identification of high-risk populations and implementation of proactive treatment strategies are crucial for improving patient prognosis. As big data and machine learning advance, the construction of visual statistical models, such as nomograms, has been extensively applied in the prognostic prediction of various cancers [12-14]. However, studies specifically focused on predicting LNM in EC remain relatively scarce, with a limited sample size [15].

This study is dedicated to building an innovative EC LNM risk prediction model that integrates multiple clinical factors and is optimized using advanced machine learning algorithms. Unlike previous studies, we not only take traditional clinical and pathological characteristics into account but also introduce a systematic integration approach based on multivariate analysis to enhance prediction accuracy (ACC) and the capacity to guide individualized treatment. Through this approach, this study offers a new perspective for LNM risk assessment and provides robust support for precise treatment and preoperative decision-making in EC.

Methods and materials

Research design

This retrospective cohort study included 541 cases of EC diagnosed and treated at The First Affiliated Hospital, Jinan University between January 2017 and January 2022. For model development and validation, the samples were randomly assigned to a training set (n = 378) and a validation set (n = 163) in a 7:3 ratio. The training set served for model building and optimization, while the validation set served to assess the models' predictive performance and robustness. This grouping approach ensures adequate training during the model development and provides a reliable independent evaluation during validation, thereby ensuring the external validity of the model. This study was approved by The First Affiliated Hospital, Jinan University.

Inclusion and exclusion criteria

Inclusion criteria: 1. Patients pathologically diagnosed with EC, including endometrioid adenocarcinoma and serous, mucinous, clear-cell, and undifferentiated carcinomas, confirmed by diagnostic curettage; 2. Patients who underwent comprehensive staging surgery and were confirmed as EC by postoperative pathological examination; 3. Patients with complete general and pathological information.

Exclusion criteria: 1. Patients who initially underwent total hysterectomy and had a second operation after EC was pathologically confirmed; 2. Patients with pre-operative radiotherapy, chemotherapy, or hormonal therapy; 3. Patients with other concurrent malignant neoplastic diseases; 4. Patients with recent systemic or local infectious lesions; 5. Patients who had received antibiotic or antiviral treatment within two weeks prior to the surgery; 6. Patients with hematological disorders or autoimmune diseases prior to the operation.

Data collection

Clinical data, including basic demographic information, pathological types, and staging of patients, were collected. The main research variables included age, body mass index (BMI), pathological grading, myometrial invasion (MI), lymphatic vascular space invasion (LVSI), sta-

tus of estrogen receptor (ER) and progesterone receptor (PR) status, tumor size, and LNM.

Machine-learning model building

The machine learning algorithms used for model development included Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and RandomForest. To guarantee the models' optimal performance, grid search cross-validation (CV) was employed for selecting the optimal hyper-parameters for each algorithm. This process entailed iterative adjustment of parameters to identify the optimal combination, thus maximizing prediction ACC and minimizing the risk of over-fitting. For the evaluation of the models' predictive capacity, a 10-fold CV was carried out on both the training and the validation sets. Using this technique, the data were partitioned into ten subsets, nine of which were utilized for model training and the rest for performance evaluation. A robust assessment of the models' generalizability was obtained by repeating this process across subsets.

Model assessment

The models' performance was evaluated using several approaches, including the receiver operating characteristic (ROC) curve, the area under the curve (AUC), sensitivity (SEN), specificity (SPE), ACC, and precision. The ROC curve visually represents the trade-off between the true (SEN) and false positive rates (1-SPE) across classification thresholds. The AUC stands for the model's overall discriminatory ability, while SEN, SPE, ACC, and precision assess the model's performance from various evaluation dimensions. Additionally, decision curve analysis (DCA) enables a comparison of predictive performance and potential practical utility among different model by considering the actual decision-making risks and the selection of prediction probability thresholds. Calibration curves were also used to assess the models' predictive ability and the consistency with actual outcomes.

Outcome measurements

1. Analysis of characteristics of LNM patients: Disparities in characteristics between LNM and non-LNM groups in both the training and validation sets were compared. 2. Identification of independent risk factors for LNM: Independent risk factors influencing LNM were identified through multivariate Logistic regression analysis. 3. Feature selection and optimization: Feature selection and model optimization were conducted using LASSO, XGBoost, RandomForest, and SVM models. 4. Predictive performance evaluation: The predictive performance of each machine-learning model was appraised through the ROC curve, AUC, SEN, SPE, and other relevant metrics. 5. Calibration capacity and consistency evaluation: The consistency of the model-predicted values with actual outcomes was examined using the calibration curves. 6. Clinical practicality analysis: The clinical applicability of each model was assessed through DCA.

Statistical analyses

The statistical analysis for this study was conducted in two parts. First, SPSS 26.0 was utilized for basic data statistical analyses, employing t-tests for normally distributed continuous variables (expressed as mean ± standard deviation) and rank-sum tests for non-normally distributed continuous variables (expressed as quartiles). Categorical variables were analyzed using chi-square tests, and multivariate Logistic regression analysis was conducted to determine independent factors associated with LNM. Further advanced analysis was performed using R language (version 4.3.2). For LASSO regression analysis, the glmnet package [16] was utilized. The XGBoost package [17] was implemented for the XGBoost model. The RandomForest model was constructed by means of the randomForest package [18], and for the SVM model, the e1071 package [19] was utilized. The models' performance was evaluated through the employment of ROC curves and AUCs. The ggplot2 package was employed for the tasks associated with graph plotting. Additionally, the cowplot package was utilized to adjust the format and optimize the layout of the graphs, enabling clear comparison of the models. For DCA, the rmda package was employed to evaluate the models' clinical applicability and the rms package to generate the calibration curve for assessing the consistency between the predicted probabilities and the actual observed results. A *P*-value less than 0.05 signified the presence of a statistical difference.

Results

Comparison of baseline data between training and validation sets

No significant differences were found between the training (n = 378) and validation (n = 163) sets in terms of age (P = 0.606), BMI (P = 0.971), prevalence of diabetes (P = 0.411), hypertension (P = 0.130), menopausal status (P = 0.884), FIGO staging (P = 0.939), pathological grading (P = 0.346), depth of MI (P = 0.544), LVSI (P = 0.800), ER status (P = 0.504), PR status (P = 0.458), tumor size (P = 0.718), pathological type (P = 0.479), and LNM (P = 0.679). Refer to **Table 1** for detailed data.

Comparison of baseline data between LNM and non-LNM patients in both training and validation sets

As indicated by **Table 2**, significant differences were observed between LNM and non-LNM groups in the training set in terms of FIGO staging (P < 0.001), pathological grading (P < 0.001), depth of MI (P < 0.001), LVSI (P < 0.001), ER (P = 0.001), PR (P = 0.001), tumor size (P = 0.010), and pathological type (P = 0.001). In the validation set, these variables also exhibited significant differences between LNM and non-LNM groups (all P < 0.05).

Multivariate logistic regression analysis results

The results (**Table 3**) demonstrated that FIGO staging (OR = 6.452, 95% CI: 2.555-16.292, P < 0.001), pathological grade (OR = 0.04, 95% CI: 0.016-0.097, P < 0.001), depth of MI (OR = 13.169, 95% CI: 3.646-47.561, P < 0.001), LVSI (OR = 4.821, 95% CI: 1.682-13.818, P = 0.003), and pathological type (OR = 0.259, 95% CI: 0.095-0.709, P = 0.009) were significant independent factors influencing LNM, while ER, PR, and tumor size did not show significant correlation with LNM (all P > 0.05).

Feature selection and optimization of machine-learning models

Based on the five selected feature factors, four machine-learning models - LASSO, XGBoost, RandomForest, and SVM - were constructed. The LASSO model was optimized by selecting the optimal penalty parameter (λ) using cross-validation (**Figure 1A**, **1B**). In the XGBoost model, increasing the number of features reduced the error rate, achieving an optimal point at a specific feature count (**Figure 1C**). For the RandomForest model, error rates varied with the number of iterations and trees, helping identify optimal parameters (**Figure 1D**, **1E**). All models incorporated the five feature factors, except SVM, which excluded pathological type (**Figure 2**).

Comparison of ROC curves in training and validation sets

The XGBoost model offered the best performance in the training set, with an AUC of 0.876 (95% CI: 0.825-0.926), relatively high SPE (85.57%) and SEN (78.08%), and an overall ACC of 84.13%. The LASSO and RandomForest models showed similar performance, with AUCs of 0.858 and 0.853, respectively, an identical SPE of 86.89%, a SEN of 78.08%, and an overall ACC of 85.19%. Although the SVM model had the lowest AUC of 0.732 (95% CI: 0.647-0.817), it had the highest SPE of 93.11% and an overall ACC of 87.30% (**Table 4**; **Figure 3A**).

The XGBoost model still performed outstandingly in the validation set, with an AUC of 0.832 (95% CI: 0.743-0.920), a SPE of 86.05%, a SEN of 70.59%, and an overall ACC of 82.82%. The LASSO model showed an AUC of 0.81 (95% CI: 0.722-0.897), a SPE of 88.37%, a SEN of 67.65%, and an overall ACC of 84.05%. The AUC of the RandomForest model was 0.812 (95% CI: 0.725-0.899), the SPE was 86.05%, the SEN was 70.59%, and the overall ACC was 82.82%. The SVM model achieved an AUC of 0.701 (95% CI: 0.577-0.826) in the validation set but still maintained the highest SPE (96.90%) and an overall ACC of 87.73% (**Table 5**; **Figure 3B**).

Calibration and decision curve analysis (DCA)

In both the training and validation sets, the calibration curves indicated that the performances of LASSO, XGBoost, RandomForest, and SVM models were close to the ideal calibration curve, suggesting a favorable consistency between the predicted and actual probabilities (**Figure 4**). Notably, the LASSO and XGBoost models demonstrated the lowest mean absolute error (MAE), with a value of 0.017 for the former and 0.027 for the latter, further corroborating their reliability in model calibration. In the DCA, the

Table 1. Comparison of baseline data between training set and validation set (7:3)

Variable	Training set (n = 378)	Validation set (n = 163)	*p*_value
Age	53.97±9.92	53.50±9.68	0.606
Body mass index	24.07±2.21	24.06±2.06	0.971
Diabetes	χ^2	χ^2	
Without	309	138	0.411
With	69	25	
Hypertension			
Without	286	133	0.130
With	92	30	
Pausimenia			
No	102	43	0.884
Yes	276	120	
FIGO staging			
I	28	11	0.939
II	125	58	
III	126	54	
IV	99	40	
Pathological grading			
G1	143	60	0.346
G2	117	60	
G3	118	43	
Myometrial invasion			
< 1/2	113	53	0.544
≥ 1/2	265	110	
LVSI			
Without	331	144	0.800
With	47	19	
Estrogen receptor			
Negative	39	20	0.504
Positive	339	143	
Progesterone receptor			
Negative	53	19	0.458
Positive	325	144	
Tumor size			
< 2 cm	143	59	0.718
≥ 2 cm	235	104	
Pathological type			
Others	28	15	0.479
Endometrioid adenocarcinoma	350	148	
LNM			
Without	305	129	0.679
With	73	34	

Note: FIGO: International Federation of Gynecology and Obstetrics; LVSI: lymphatic vascular space invasion; LNM, lymph node metastasis.

LASSO, XGBoost, and RandomForest models demonstrated relatively high net benefits across many threshold ranges, particularly in the validation set, suggesting their greater potential value in clinical applications. Although the SVM model performed relatively less effectively, it still retained certain application value within specific threshold ranges (**Figure 5**).

Table 2. Comparison of baseline data between LNM and non-LNM patients in both training and validation sets

Variable	Training set			Validation set		
	Non-LNM group (n = 305)	LNM group (n = 73)	*p*_value	Non-LNM group (n = 129)	LNM group (n = 34)	*p*_value
Age	54.05±10.15	53.63±8.96	0.727	53.55±9.70	53.29±9.74	0.892
Body mass index	24.02±2.13	24.25±2.54	0.486	24.02±2.10	24.22±1.94	0.605
Diabetes						
Without	244	65	0.072	112	26	0.136
With	61	8		17	8	
Hypertension						
Without	228	58	0.401	106	27	0.712
With	77	15		23	7	
Pausimenia						
No	82	20	0.929	32	11	0.374
Yes	223	53		97	23	
FIGO staging						
I	19	9	< 0.001	6	5	< 0.001
II	116	9		53	5	
III	105	21		45	9	
IV	65	34		25	15	
Pathological grading						
G1	124	19	< 0.001	53	7	0.047
G2	79	38		42	18	
G3	102	16		34	9	
Myometrial invasion						
< 1/2	108	5	< 0.001	49	4	0.004
≥ 1/2	197	68		80	30	
LVSI						
Without	284	47	< 0.001	121	23	< 0.001
With	21	26		8	11	
Estrogen receptor						
Negative	24	15	0.001	11	9	0.005
Positive	281	58		118	25	
Progesterone receptor						
Negative	34	19	0.001	11	8	0.015
Positive	271	54		118	26	
Tumor size						
< 2 cm	125	18	0.010	51	8	0.084
≥ 2 cm	180	55		78	26	
Pathological type						
Others	16	12	0.001	7	8	0.001
Endometrioid adenocarcinoma	289	61		122	26	

Note: FIGO: International Federation of Gynecology and Obstetrics; LVSI: lymphatic vascular space invasion; LNM: lymph node metastasis.

Nomogram based on five feature factors

The nomogram serves to calculate the total risk score for individual patients and predict the probability of a specific clinical outcome (such as disease recurrence or metastasis). In the figure, each feature factor corresponds to an independent scale line. Based on the scale line

Table 3. Multivariate logistic regression analysis

Variable	β	Standard error	χ^2	P	OR	Lower bound	Upper bound
FIGO staging	1.864	0.473	15.565	< 0.001	6.452	2.555	16.292
Pathological grading	-3.225	0.454	50.482	< 0.001	0.04	0.016	0.097
Myometrial invasion	2.578	0.655	15.481	< 0.001	13.169	3.646	47.561
LVSI	1.573	0.537	8.57	0.003	4.821	1.682	13.818
Estrogen receptor	0.562	0.616	0.833	0.362	1.754	0.525	5.867
Progesterone receptor	-0.196	0.554	0.125	0.723	0.822	0.277	2.435
Tumor size	0.657	0.43	2.339	0.126	1.929	0.831	4.478
Pathological type	-1.349	0.513	6.923	0.009	0.259	0.095	0.709

Note: FIGO: International Federation of Gynecology and Obstetrics; LVSI: lymphatic vascular space invasion.

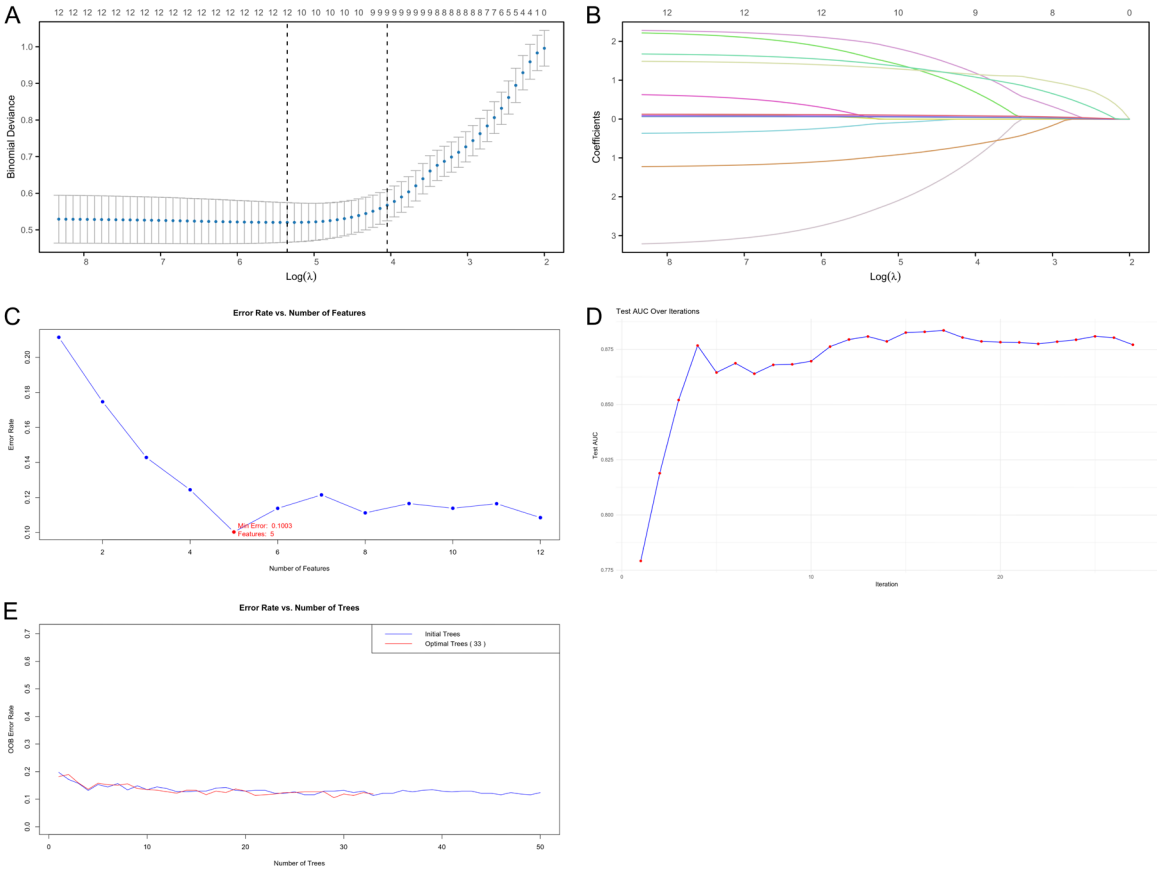

Figure 1. Feature factor selection in four machine-learning models. A. Path diagram of penalty parameter selection in the LASSO model. This figure shows the model's variation of mean squared error (MSE) with the change of regularization parameter (λ). The vertical dashed line indicates the optimal λ determined via cross-validation. B. Coefficient path diagram in the LASSO model. This graph shows how the coefficients of different features change at different λ values, helping to determine which variables are retained during regularization. C. Influence of different feature quantities on the error rate in the XGBoost model. This figure illustrates how the model's error rate varies as the selected feature number elevates and indicates the optimal feature number (i.e. the point with the minimum error rate). D. Variation of total error with the number of iterations in the RandomForest model. The red and blue lines in the figure represent the error rates of the training and validation sets, respectively, demonstrating the model's performance as the number of iterations increases. E. Relationship between the number of trees and the error rate in the RandomForest model. This graph presents the trend of error rate variation as the number of trees in the RandomForest increases in both the training and validation sets. Note: LASSO: Least Absolute Shrinkage and Selection Operator; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

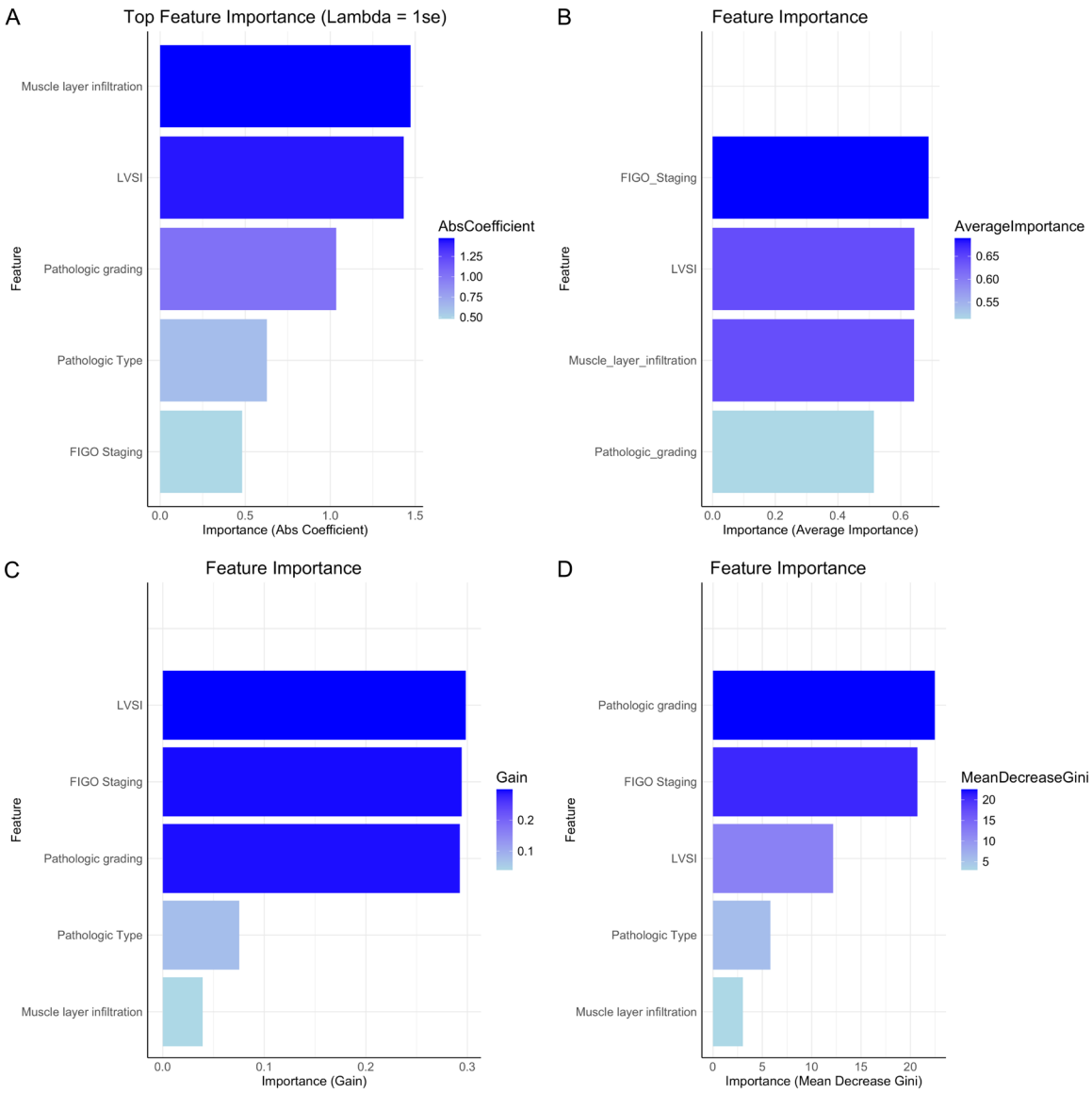

Figure 2. Incorporation of feature factors in machine-learning models. A. Incorporation of feature factors in the LASSO model. B. Incorporation of feature factors in the SVM model. C. Incorporation of feature factors in the XGBoost model. D. Incorporation of feature factors in the RandomForest model. Note: LASSO: Least Absolute Shrinkage and Selection Operator; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

corresponding to each feature factor, clinicians can determine the specific score for a patient based on their condition. These scores are summed to yield the total score, which corresponds to the probability line at the bottom of the nomogram, showing the predicted probability of the patient experiencing this clinical outcome. For example, the higher the scores of FIGO staging, pathological grading, and MI, the greater the risk for the patient. In this way, the nomogram can assist clinicians in making more precise risk assessments and treatment decisions for individual patients (**Figure 6**).

Discussion

In this study, we successfully built and optimized several machine-learning models for predicting lymph node metastasis (LNM) risk in endometrial cancer (EC). By analyzing clinicopathological data and employing multivariate Logistic regression, we identified independent risk factors strongly linked to LNM. Four machine-learning algorithms, namely the LASSO, XGBoost, RandomForest, and SVM, were utilized for feature selection and model construction. Of them, the XGBoost model outperformed

Table 4. ROC curve parameters of models in the training set

Marker	AUC	CI_lower_upper	Specificity	Sensitivity	Youden_index	Cut_off	Accuracy	Precision	F1_Score
LASSO	0.858	0.809-0.906	86.89%	78.08%	64.97%	0.226	85.19%	78.08%	67.06%
XGBoost	0.876	0.825-0.926	85.57%	78.08%	63.66%	0.473	84.13%	78.08%	65.52%
RandomForest	0.853	0.799-0.906	86.89%	78.08%	64.97%	0.104	85.19%	78.08%	67.06%
SMV	0.732	0.647-0.817	93.11%	63.01%	56.13%	0.124	87.30%	63.01%	65.71%

Note: AUC: area under the curve; LASSO: Least Absolute Shrinkage and Selection Operator; XGBoost: Extreme Gradient Boosting; SVM: Support Vector Machine.

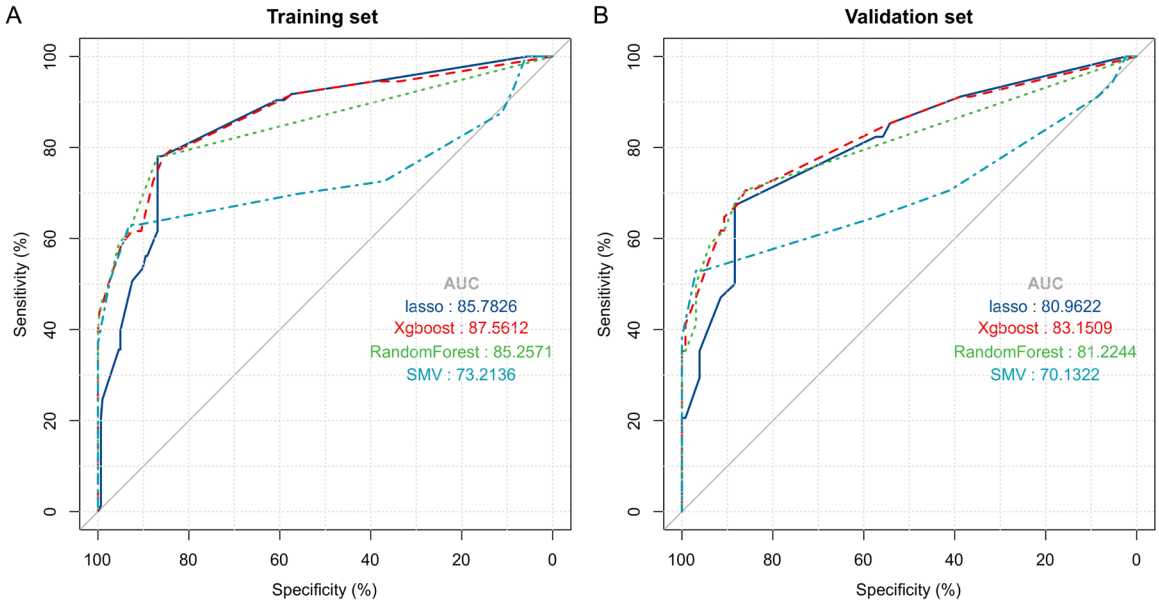

Figure 3. Comparison of ROC curves of each model in training and validation sets. A. ROC curves in the training set, demonstrating the classification performance of various machine learning models, including LASSO, XGBoost, RandomForest, and SVM models. B. ROC curves in the validation set, depicting the classification performance of LASSO, XGBoost, RandomForest, and SVM models. Note: LASSO: Least Absolute Shrinkage and Selection Operator; XGBoost: Extreme Gradient Boosting; SVM: Support Vector Machine.

Table 5. ROC curve parameters of models in the validation set

Marker	AUC	CI_lower_upper	Specificity	Sensitivity	Youden_index	Cut_off	Accuracy	Precision	F1_Score
LASSO	0.81	0.722-0.897	88.37%	67.65%	56.02%	0.226	84.05%	67.65%	63.89%
XGBoost	0.832	0.743-0.920	86.05%	70.59%	56.63%	0.464	82.82%	70.59%	63.16%
RandomForest	0.812	0.725-0.899	86.05%	70.59%	56.63%	0.021	82.82%	70.59%	63.16%
SMV	0.701	0.577-0.826	96.90%	52.94%	49.84%	0.124	87.73%	52.94%	64.29%

Note: ROC: receiver operating characteristic; LASSO: Least Absolute Shrinkage and Selection Operator; XGBoost: Extreme Gradient Boosting; SVM: Support Vector Machine.

others across all performance metrics, with an AUC of 0.876 in the training set and 0.832 in the validation set, demonstrating its high discriminative ability and predictive ACC. Li et al. [20] found that LVSI, deep MI, and CA125 levels were intimately associated with LNM in EC patients. Sari et al. [21] also identified LVSI and pelvic LNM as independent risk factors for paraaortic LNM in EC. Additionally, calibration curve analysis and DCA further validated the predictive ability and clinical application value of the XGBoost model. Within most of the risk threshold ranges, the XGBoost model exhibited relatively high net benefits, suggesting its potential application value in clinical decision-making support.

In comparison to previous studies, our research not only adopted multiple machine-learning algorithms but also enhanced the models' pre-

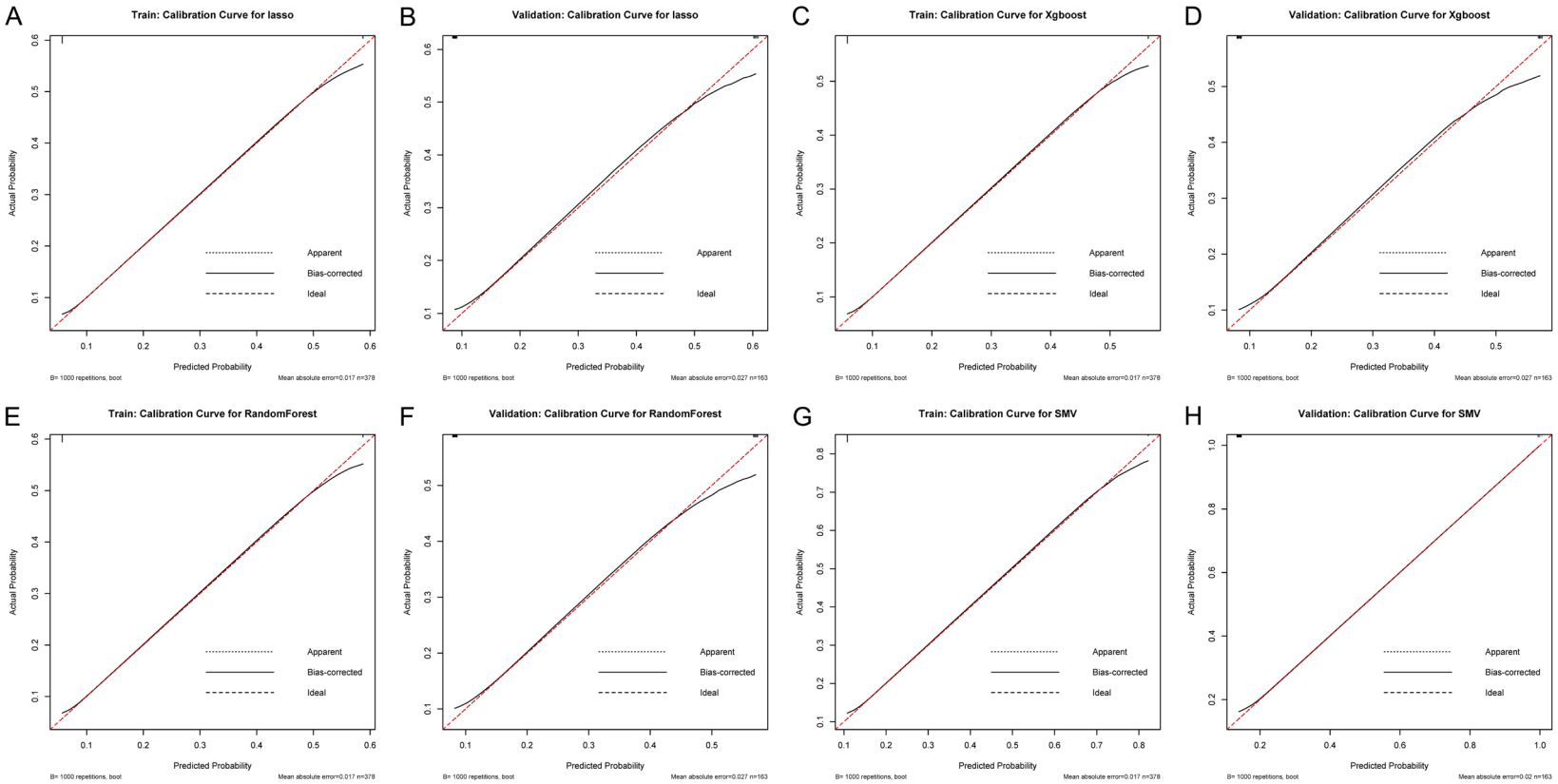

Figure 4. Calibration curves of the four models in training and validation sets. A, B. Calibration curves of the LASSO model in the training set and the validation set, respectively, illustrating the fitting between the predicted probabilities and the actual probabilities. C, D. Calibration curves of the XGBoost model in the training set and the validation set, respectively, demonstrating the fitting between the predicted probabilities and the actual probabilities. E, F. Calibration curves of the RandomForest model in the training set and the validation set, respectively, depicting the fitting between the predicted probabilities and the actual probabilities. G, H. Calibration curves of the SVM model in the training set and the validation set, respectively, presenting the fitting between the predicted probabilities and the actual probabilities.

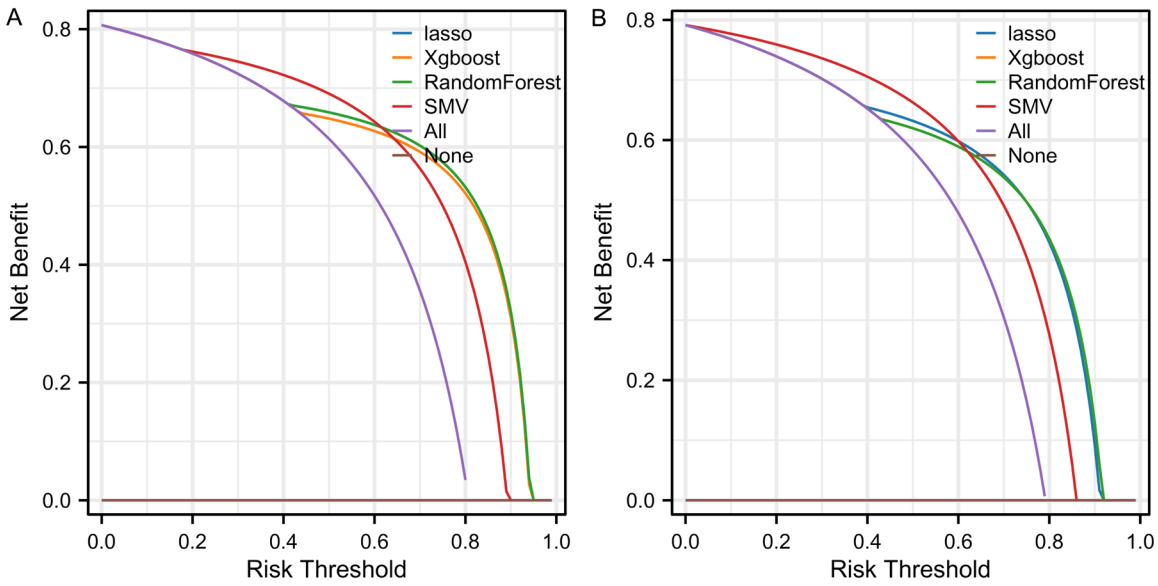

Figure 5. DCA of the four models in training and validation sets. A. Decision curves in the training set, showing the net benefits of the LASSO, XGBoost, RandomForest, and SVM models under different probability thresholds. B. Decision curves in the validation set, showing the net benefits of the LASSO, XGBoost, RandomForest, and SVM models under different probability thresholds. Note: DCA: decision curve analysis; LASSO: Least Absolute Shrinkage and Selection Operator; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.

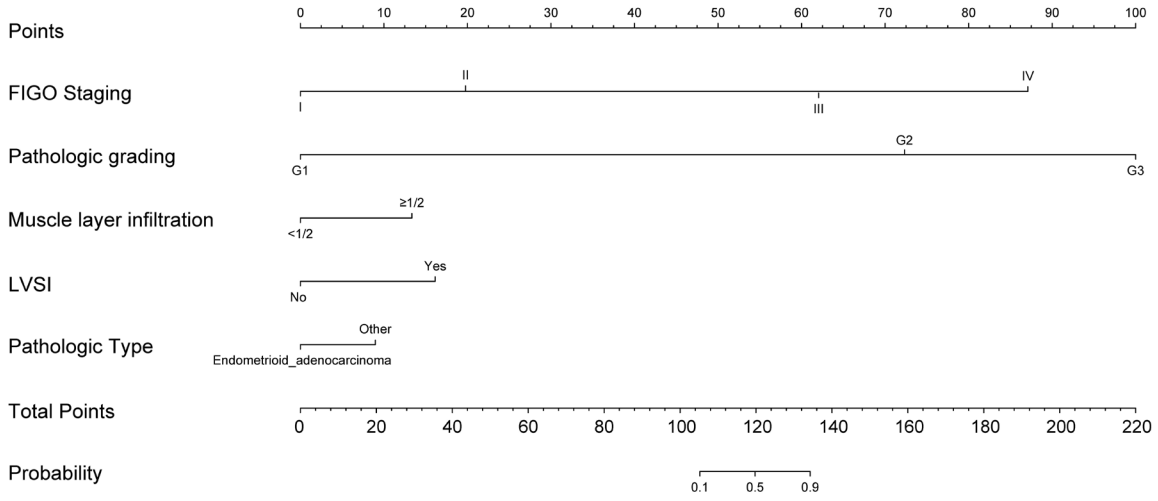

Figure 6. A nomogram based on five feature factors.

dictive performance through systematic feature selection and model optimization. Particularly, the XGBoost model outperformed other models in multiple indicators such as AUC, SPE, and SEN. This is consistent with findings from a systematic review and meta-analysis by Ren et al. [22], which indicated that machine-learning models combining clinical and radiomics features had higher predictive efficiency in predicting LNM. While our results align with this trend, discrepancies in other studies may be attributed to differences in sample sizes, study designs, feature selection processes, and model-tuning strategies. Furthermore, the specific variables selected in this study, such as FIGO staging, pathological grading, and depth of MI, demonstrated significant roles in predicting LNM, with their predictive implications being widely acknowledged in other related studies. For instance, Sun et al. [23] identified positive

peritoneal cytology, cervical stromal infiltration, MI ≥ 1/2, and LVSI as independent risk factors for pelvic LNM in non-endometrioid adenocarcinoma patients. Besides, Zhang et al. [24] pointed out that EC patients with a tumor diameter of ≥ 2 cm, a BMI of ≥ 24 kg/m^2, low-grade differentiation, and cervical stromal invasion had an elevated risk of sentinel LNM. In summary, our study reaffirms the effectiveness of advanced feature selection and model optimization to improve LNM prediction accuracy, with the XGBoost achieving the most robust performance.

In this study, the XGBoost, LASSO, and RandomForest models all demonstrated relatively high predictive capabilities in assessing the LNM risk in EC patients. However, the XGBoost model significantly outperformed the other models in terms of the AUC and other evaluation indicators. Specifically, the AUC was 0.876 in the training set and 0.832 in the validation set in the XGBoost model, indicating a relatively high discriminative ability and stability. Moreover, the XGBoost model exhibited excellent performance in terms of SPE and SEN, further validating its applicability and robustness across different data sets. In contrast, while the LASSO and RandomForest models also showed solid predictive performance, their AUCs and other indicators were slightly lower compared to those of the XGBoost model. Yang et al. [25] also showed that, when combining radiomics features and clinical characteristics, the XGBoost model had superior performance in predicting LNM in EC. Overall, XGBoost was confirmed as the optimal model for accurate and stable LNM prediction.

Machine-learning algorithms, especially XGBoost, excel at handling high-dimensional data, automatically processing missing values, and selecting relevant features. These capabilities allows XGBoost to better capture complex relationships in the data and effectively avoid overfitting during model building [26, 27]. Compared with traditional statistical methods, machine-learning algorithms can be trained through multiple iterations to optimize model parameters, thereby significantly enhancing prediction ACC and stability [28]. Moreover, XGBoost's ability to handle nonlinear relationships is particularly valuable in medical predictive models, as the relationships between many biometrics and clinical features are often complex and nonlinear [29]. The study by Miller et al. [30] highlighted the potential of machine-learning algorithms to improve predictive performance when combining clinical and molecular features. These advantages highlight XGBoost's strong potential in managing complex medical data for robust predictive modeling.

In the feature-screening procedure of this study, FIGO staging, pathological grading, depth of MI, LVSI, and pathological type emerged as critical factors in predicting LNM. Multivariate Logistic regression analysis demonstrated that these variables were strongly correlated with the LNM risk and were uniformly selected as crucial features by the XGBoost, LASSO, and RandomForest models. These variables have also been extensively verified to possess important predictive values in other investigations. For instances, Ueno et al. [31] revealed that LVSI, pathological grading, and tumor size played an important role in predicting LNM in EC, which is in accordance with our findings. In addition, Schivardi et al. [32] found that the combination of molecular typing and pathological features could significantly enhance the predictive capacity for the recurrence risk of EC patients, particularly in cases with LNM. Our study further validates the importance of these variables in accurately predicting LNM risk.

The significance of these variables in our models may be attributed to their direct reflection of tumor aggressiveness and metastatic potential. LVSI has been widely recognized as an important predictor of LNM as a measure of tumor invasion through blood vessels and lymphatic vessels [33]. The depth of MI reflects the spread of the tumor within the uterine wall, and deep MI tends to predict a higher risk of metastasis [34]. Pathological types and pathological grading reveal the biological behavior and malignancy degree of the tumor, which have been proven to be strongly related to patient outcomes in multiple studies. Huang et al. [35] constructed a combined ratio model by analyzing ERα, PR, P53, and Ki67, which showed significant ACC in predicting the LNM risk in low-risk EC patients, further supporting the findings of this study. Zanfagnin et al. [36] also noted that LVSI, gross pelvic LNM, and uterine serous carcinoma, which were also identified by this study to be key variables in predicting LNM in

EC, were closely linked to the occurrence of multiple LNM. Collectively, these factors enhance the model's capacity to reflect tumor characteristics and aid in predicting LNM with greater precision.

Limitations and prospects

Despite the construction of an effective LNM risk prediction model via multiple machine-learning algorithms in this study, several limitations remain. First of all, the relatively limited sample size and the data being derived from a single center may potentially affect the model's generalizability and its applicability across diverse populations. The absence of external validation also constrains the model's performance in other independent data sets. Additionally, variations in ethnicity and clinical backgrounds might exert an influence on the model's predictive performance. Future research should validate the model's robustness using multi-center data and introduce novel clinical or molecular-biological features for further model optimization, thereby broadening its application potential in other cancer types.

Conclusion

This study successfully constructed EC LNM risk prediction models using multiple machine-learning algorithms, and XGBoost demonstrated significant potential in clinical applications. By integrating a substantial amount of clinical and pathological data, the XGBoost model can provide valuable support for clinicians in accurate risk evaluation and decision-making. This approach has the potential to improve patient prognosis and treatment outcomes by enabling more precise, individualized management of EC patients at risk for LNM.

Disclosure of conflict of interest

None.

Address correspondence to: Yuxia Wang, Department of Reproductive Medicine, The First Affiliated Hospital, Jinan University, Guangzhou 510000, Guangdong, China. Tel: +86-020-38688888; E-mail: sunneywang@126.com

References

[1] Crosbie EJ, Kitson SJ, McAlpine JN, Mukhopadhyay A, Powell ME and Singh N. Endometrial cancer. Lancet 2022; 399: 1412-1428.

[2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A and Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71: 209-249.

[3] Sun KX, Zheng RS, Zuo J, Zhang SW, Zeng HM, Wang SM, Chen R, Li L, Wei WW and He J. The incidence and mortality of endometrial cancer in China, 2015. Zhonghua Yi Xue Za Zhi 2022; 102: 1987-1992.

[4] Sun S, Zou L, Wang T, Liu Z, He J, Sun X, Zhong W, Zhao F, Li X, Li S, Zhu H, Ma Z, Wang W, Jin M, Zhang F, Hou X, Wei L and Hu K. Effect of age as a continuous variable in early-stage endometrial carcinoma: a multi-institutional analysis in China. Aging (Albany NY) 2021; 13: 19561-19574.

[5] Morrison CW, Sanjasaz KN, Nathanson SD, Raina-Hukku S, Pinkney DM and Davenport AA. Dedifferentiated endometrial carcinoma metastasis to axillary lymph node: a case report. J Med Case Rep 2023; 17: 451.

[6] Baum S, Alkatout I, Proppe L, Kotanidis C, Rody A, Laganà AS, Sommer S and Gitas G. Surgical treatment of endometrioid endometrial carcinoma - laparotomy versus laparoscopy. J Turk Ger Gynecol Assoc 2022; 23: 233-240.

[7] Cabrera S, Bebia V, López-Gil C, Luzarraga-Aznar A, Denizli M, Salazar-Huayna L, Abdessayed N, Castellví J, Colas E and Gil-Moreno A. Molecular classification improves preoperative risk assessment of endometrial cancer. Gynecol Oncol 2024; 189: 56-63.

[8] Liu M and Peng J. A meta-analysis of the effect of pelvic and para-aortic lymph node dissection on the prognosis of patients with endometrial cancer. Biotechnol Genet Eng Rev 2024; 40: 2926-2944.

[9] Miyamoto M, Hada T, Ishibashi H, Iwahashi H, Kakimoto S, Suzuki R, Sakamoto T, Matsuura H, Tsuda H and Takano M. A new model to improve the prediction of prognosis of endometrial carcinoma by combining traditional classification with the presence of tumor-infiltrating lymphocytes. Anticancer Res 2021; 41: 1047-1053.

[10] Matsumoto YK, Himoto Y, Nishio M, Kikkawa N, Otani S, Ito K, Yamanoi K, Kato T, Fujimoto K, Kurata Y, Moribata Y, Yoshida H, Minamiguchi S, Mandai M, Kido A and Nakamoto Y. Nodal infiltration in endometrial cancer: a prediction model using best subset regression. Eur Radiol 2024; 34: 3375-3384.

[11] Shazly SA, Coronado PJ, Yılmaz E, Melekoglu R, Sahin H, Giannella L, Ciavattini A, Carpini GD, Di Giuseppe J, Yordanov A, Karakadieva K, Nedelcheva NM, Vasileva-Slaveva M, Alcazar JL, Chacon E, Manzour N, Vara J, Karaman E, Karaaslan O, Hacıoğlu L, Korkmaz D, Onal C,

Knez J, Ferrari F, Hosni EM, Mahmoud ME, Elassall GM, Abdo MS, Mohamed YI and Abdelbadie AS; Middle-Eastern College of Obstetricians and Gynaecologists (MCOG) Muti-Center Studies (MCS) office and Artificial Intelligence Unit (AI). Endometrial cancer individualized scoring system (ECISS): a machine learning-based prediction model of endometrial cancer prognosis. Int J Gynaecol Obstet 2023; 161: 760-768.

[12] Jin F, Liu W, Qiao X, Shi J, Xin R and Jia HQ. Nomogram prediction model of postoperative pneumonia in patients with lung cancer: a retrospective cohort study. Front Oncol 2023; 13: 1114302.

[13] Chen L, Ma X, Dong H, Qu B, Yang T, Xu M, Sheng G, Hu J and Liu A. Construction and assessment of a joint prediction model and nomogram for colorectal cancer. J Gastrointest Oncol 2022; 13: 2406-2414.

[14] Lu Z, Sun J, Wang M, Jiang H, Chen G and Zhang W. A nomogram prediction model based on clinicopathological combined radiological features for metachronous liver metastasis of colorectal cancer. J Cancer 2024; 15: 916-925.

[15] Zhang J, Wang D, Peng L, Shi X, Shi Y and Zhang G. Preoperative evaluation and a nomogram prediction model for pelvic lymph node metastasis in endometrial cancer. Eur J Surg Oncol 2024; 50: 108230.

[16] Engebretsen S and Bohlin J. Statistical predictions with glmnet. Clin Epigenetics 2019; 11: 123.

[17] Li K, Yao S, Zhang Z, Cao B, Wilson CM, Kalos D, Kuan PF, Zhu R and Wang X. Efficient gradient boosting for prognostic biomarker discovery. Bioinformatics 2022; 38: 1631-1638.

[18] Jiang H, Chen H, Wang Y and Qian Y. Novel molecular subtyping scheme based on in silico analysis of cuproptosis regulator gene patterns optimizes survival prediction and treatment of hepatocellular Carcinoma. J Clin Med 2023; 12: 5767.

[19] Yang L, Pan X, Zhang Y, Zhao D, Wang L, Yuan G, Zhou C, Li T and Li W. Bioinformatics analysis to screen for genes related to myocardial infarction. Front Genet 2022; 13: 990888.

[20] Li Y, Cong P, Wang P, Peng C, Liu M and Sun G. Risk factors for pelvic lymph node metastasis in endometrial cancer. Arch Gynecol Obstet 2019; 300: 1007-1013.

[21] Sari ME, Yalcin İ, Sahin H, Meydanli MM and Gungor T. Risk factors for paraaortic lymph node metastasis in endometrial cancer. Int J Clin Oncol 2017; 22: 937-944.

[22] Ren Z, Chen B, Hong C, Yuan J, Deng J, Chen Y, Ye J and Li Y. The value of machine learning in preoperative identification of lymph node metastasis status in endometrial cancer: a systematic review and meta-analysis. Front Oncol 2023; 13: 1289050.

[23] Sun Y, Wang Y, Cheng X, Wu W, Liu Q, Chen X and Ren F. Risk factors for pelvic and paraaortic lymph node metastasis in non-endometrioid endometrial cancer. Eur J Surg Oncol 2024; 50: 108260.

[24] Zhang Y, Liu H, Han X and Tang Y. Analysis of risk factors for sentinel lymph node metastasis in patients with endometrial cancer. Am J Transl Res 2022; 14: 8650-8658.

[25] Yang LY, Siow TY, Lin YC, Wu RC, Lu HY, Chiang HJ, Ho CY, Huang YT, Huang YL, Pan YB, Chao A, Lai CH and Lin G. Computer-aided segmentation and machine learning of integrated clinical and diffusion-weighted imaging parameters for predicting lymph node metastasis in endometrial cancer. Cancers (Basel) 2021; 13: 1406.

[26] Noorunnahar M, Chowdhury AH and Mila FA. A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. PLoS One 2023; 18: e0283452.

[27] Montomoli J, Romeo L, Moccia S, Bernardini M, Migliorelli L, Berardini D, Donati A, Carsetti A, Bocci MG, Wendel Garcia PD, Fumeaux T, Guerci P, Schüpbach RA, Ince C, Frontoni E and Hilty MP; RISC-19-ICU Investigators. Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. J Intensive Med 2021; 1: 110-116.

[28] Su S and Wang J. Machine learning prediction of contents of oxygenated components in bio-oil using extreme gradient boosting method under different pyrolysis conditions. Bioresour Technol 2023; 379: 129040.

[29] Li B, Eisenberg N, Beaton D, Lee DS, Aljabri B, Verma R, Wijeysundera DN, Rotstein OD, de Mestral C, Mamdani M, Roche-Nagle G and Al-Omran M. Using machine learning (XGBoost) to predict outcomes after infrainguinal bypass for peripheral artery disease. Ann Surg 2024; 279: 705-713.

[30] Miller HA, Tran A, LyBarger KS and Frieboes HB. A clinical marker-based modeling framework to preoperatively predict lymph node and vascular space involvement in endometrial cancer patients. Eur J Surg Oncol 2024; 50: 107309.

[31] Ueno Y, Yoshida E, Nojiri S, Kato T, Ohtsu T, Takeshita T, Suzuki S, Yoshida H, Kato K, Itoh M, Notomi T, Usui K, Sozu T, Terao Y, Kawaji H and Kato H. Use of clinical variables for preoperative prediction of lymph node metastasis in endometrial cancer. Jpn J Clin Oncol 2024; 54: 38-46.

[32] Schivardi G, Caruso G, De Vitis LA, Cucinella G, Multinu F, Zanagnolo V, Baiocchi G, De Brot L, Occhiali T, Vizzielli G, Giuntoli R, Fought AJ, McGree ME, Shahi M, Mariani A and Glaser GE. Impact of molecular classification on recurrence risk in endometrial cancer patients with lymph node metastasis: multicenter retrospective study. Int J Gynecol Cancer 2024; 34: 1561-1569.

[33] Tutkun Kilinc EC, Korkmaz V and Yalcin HR. Factor affecting lymph node metastasis in uterine papillary serous carcinomas: a retrospective analysis. J Obstet Gynaecol 2022; 42: 3725-3730.

[34] Li Q and Zhang X. Prediction of high-risk factors for ovarian metastasis in patients with endometrial cancer: a large-sample retrospective case-control study. Int J Gynaecol Obstet 2023; 161: 144-150.

[35] Huang Y, Jiang P, Kong W, Tu Y, Li N, Wang J, Zhou Q and Yuan R. Comprehensive assessment of ERα, PR, Ki67, P53 to predict the risk of lymph node metastasis in low-risk endometrial cancer. J Invest Surg 2023; 36: 2152508.

[36] Zanfagnin V, Huang Y, Mc Gree ME, Weaver AL, Casarin J, Multinu F, Cappuccio S, Ferrero A, Mariani A and Glaser GE. Predictors of extensive lymphatic dissemination and recurrences in node-positive endometrial cancer. Gynecol Oncol 2019; 154: 480-486.