Risk Management and Healthcare Policy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Prediction Models of Microinvasive Cervical Cancer in High-Grade Squamous Intraepithelial Lesion Treatment by Loop Electrosurgical Excision Procedure

Maodan Huang[1], Xiaohong Chen[1], Xin Lin[1], Yuxiang Yang[1], Lu Liu[2], Youzhong Zhang[3], Ronglong Wang[1], Wei Chen[4]

[1]Department of Obstetrics and Gynecology, Zhangzhou Affiliated Hospital of Fujian Medical University, Zhangzhou, 363000, People's Republic of China; [2]Department of Obstetrics and Gynecology, The Second Hospital, Cheeloo College of Medicine, Shandong University, Jinan, 250033, People's Republic of China; [3]Department of Obstetrics and Gynecology, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, 250012, People's Republic of China; [4]School of Radiology, Shandong First Medical University and Shandong Academy of Medical Sciences, Tai'an, 271016, People's Republic of China

Correspondence: Ronglong Wang, Department of Obstetrics and Gynecology, Zhangzhou Affiliated Hospital of Fujian Medical University, Zhangzhou, 363000, People's Republic of China, Email earthfire1999@163.com; Wei Chen, School of Radiology, Shandong First Medical University and Shandong Academy of Medical Sciences, Tai'an, 271016, People's Republic of China, Email chenwei9320@sdfmu.edu.cn

Objective: The implementation of comprehensive microinvasive cervical cancer (MIC) risk assessment in high-grade squamous intraepithelial lesion (HSIL) patients undergoing loop electrosurgical excision procedure (LEEP) is critical to optimize treatment strategies and improve patient outcomes.

Methods: From March 2017 to January 2024, a total of 3066 eligible patients with HSIL were retrospectively enrolled from two hospitals and assigned into one training cohort (n = 2084), one internal validation cohort (579) and one external testing cohort (n = 403). Four feature selection methods (Random Forest, Lasso regression, Boruta algorithm, and Extreme Gradient Boosting) were employed to identify key predictive factors from the training cohort. Then, four machine learning models were developed and evaluated using comprehensive metrics. The optimal model was visualized through interpretable techniques and operationalized as a web-based clinical decision support system for real-world implementation.

Results: Six clinical predictive variables were identified, including surgical margins, endocervical curettage (ECC), TCT status, HPV status, Transformation Zone (TZ) type and Age. The optimal model demonstrated good predictive performance, achieving an area under the receiver operating characteristic curve (AUC) of 0.822 (95% CI: 0.793–0.852) in the internal validation cohort and 0.802 (95% CI: 0.730–0.874) in the external validation cohort.

Conclusion: The machine learning-based model can accurately assess the risk of MIC during the treatment of HSIL with LEEP, potentially aiding in the selection of appropriate treatment and surveillance strategies in clinical practice.

Keywords: high-grade squamous intraepithelial lesions, microinvasive cervical cancer, interpretable machine learning, visualization, prediction model

Introduction

Cervical cancer remains a significant global health concern, particularly in developing countries.[1] The development of prophylactic HPV vaccines marked a significant milestone in cancer prevention, targeting the primary etiological agent of cervical cancer. However, global disparities in vaccine accessibility persist, particularly in low-income countries where less than one-third have widespread HPV vaccination programs.[2] Consequently, the management of cervical squamous intraepithelial lesions remains a critical strategy in preventing cervical cancer. Cervical squamous intraepithelial lesions are categorized into low-grade squamous intraepithelial lesions (LSIL) and high-grade squamous intraepithelial lesions

Received: 24 April 2025
Accepted: 3 September 2025
Published: 6 September 2025

(HSIL), with approximately 30% of HSIL cases potentially progressing to cervical cancer.[3–5] Cervical conization is the primary treatment for HSIL, including LEEP and Cold Knife Conization.[6]

Microinvasive cervical cancer (MIC) represents an early stage of cervical malignancy, typically diagnosed through histological evaluation of biopsy specimens or cervical conization specimens.[7] The treatment of MIC lacks a unified standard, with therapeutic strategies tailored to individual patient circumstances. For women who do not wish to preserve fertility and in whom lymphovascular space invasion is absent, hysterectomy is commonly considered as a definitive treatment option.[8] The management of patients with HSIL requires careful consideration of the potential for MIC progression. Conducting a comprehensive risk assessment for MIC in HSIL patients is not only crucial for reducing the likelihood of repeated surgeries but also for minimizing patient trauma and economic burden.

In recent years, logistic regression has frequently been used to investigate risk factors for progression to invasive cervical cancer following surgical treatment of HSIL.[9,10] However, this method can overlook potential covariation or non-linear relationships among variables, reducing its effectiveness in addressing the more complex, non-linear aspects of disease progression.[11] Moreover, existing prediction tools have notable limitations, including the lack of user-friendly interfaces, inability to provide real-time risk assessments, and insufficient capacity to model complex variable relationships. These shortcomings hinder clinicians' ability to efficiently and accurately evaluate the risk of progression to MIC in HSIL patients. In addition, no publicly available online tool currently exists to facilitate user-friendly, real-time risk assessments for both clinicians and patients.

In recent years, healthcare researchers have shown growing interest in applying machine learning (ML) techniques developed using data extracted from electronic medical records (EMR). ML technologies have expanded at an unprecedented rate, driving rapid innovation in the field of healthcare. Compared with conventional statistical methods, ML algorithms impose fewer restrictions on input data and demonstrate robust capacity for modeling complex datasets, thereby fueling their increasing adoption in clinical practice. For instance, Hu et al[12] developed an explainable prediction model for acute kidney injury (AKI) in critically ill children using machine learning techniques. Similarly, Chen et al[13] developed a machine learning model to predict the risk of intraoperative hemorrhage during cesarean scar ectopic pregnancy (CSEP) surgeries. Despite the advantages of ML, the inherent complexity of these algorithms—often described as the "black box" phenomenon—makes direct interpretation challenging.[14] Consequently, model explanation techniques have become indispensable for enhancing clinical applicability and trust. Among the available methods, SHapley Additive exPlanations (SHAP) has been leveraged to elucidate how individual variables contribute to model outputs.[15] However, few investigations have specifically applied SHAP to explore the risk of progression to invasive cervical cancer among patients with HSIL, leaving a critical gap in understanding and application of predictive models in this context.

This study aims to develop and validate an interpretable ML model through a multicenter retrospective analysis for the early and accurate prediction of progression from HSIL to MIC. The study will identify feature importance and utilize the SHAP method to interpret the model, providing insights to assist clinicians in making informed treatment decisions. Additionally, the model will be deployed in a web-based application, enabling clinicians to directly use the prediction tool without requiring any programming knowledge.

Methods

Study Participants

This is a multicenter retrospective study primarily utilizing EMR data from two hospitals. The study is composed of three main stages: the selection of research subjects and screening of predictive variables, the construction and evaluation of multiple prediction models, and the development of a web-based application for the optimal model.

We retrieved clinical data of 3692 patients with a pathological diagnosis of cervical high-grade squamous intraepithelial lesions (HSIL) who underwent loop electrosurgical excision procedure (LEEP) between March 2017 and June 2024 at two institutions in China. The inclusion criteria for patients were as follows: a confirmed diagnosis of HSIL via biopsy, undergoing LEEP treatment, having clear resection margins postoperatively, and having complete clinical and treatment data. The exclusion criteria were: (a) patients aged <25 years or >65 years; (b) incomplete clinical

or pathological data; and (c) prior treatment history. The processes for inclusion and exclusion are illustrated in Figure 1. Finally, a total of 3066 patients were enrolled, comprising 2663 patients from Institution I and 403 patients from Institution II. During the modeling process, patients from Institution I were stratified into a training cohort (n=2084) and an internal validation cohort (n=579) based on their treatment timeline. Patients from Institution II were designated as the external validation cohort.

We collected detailed clinical information of the included cohort, including age, reproductive history, menopausal status, Transformation Zone (TZ) type, Thinprep cytologic test (TCT) results, HPV status, surgical margins, and

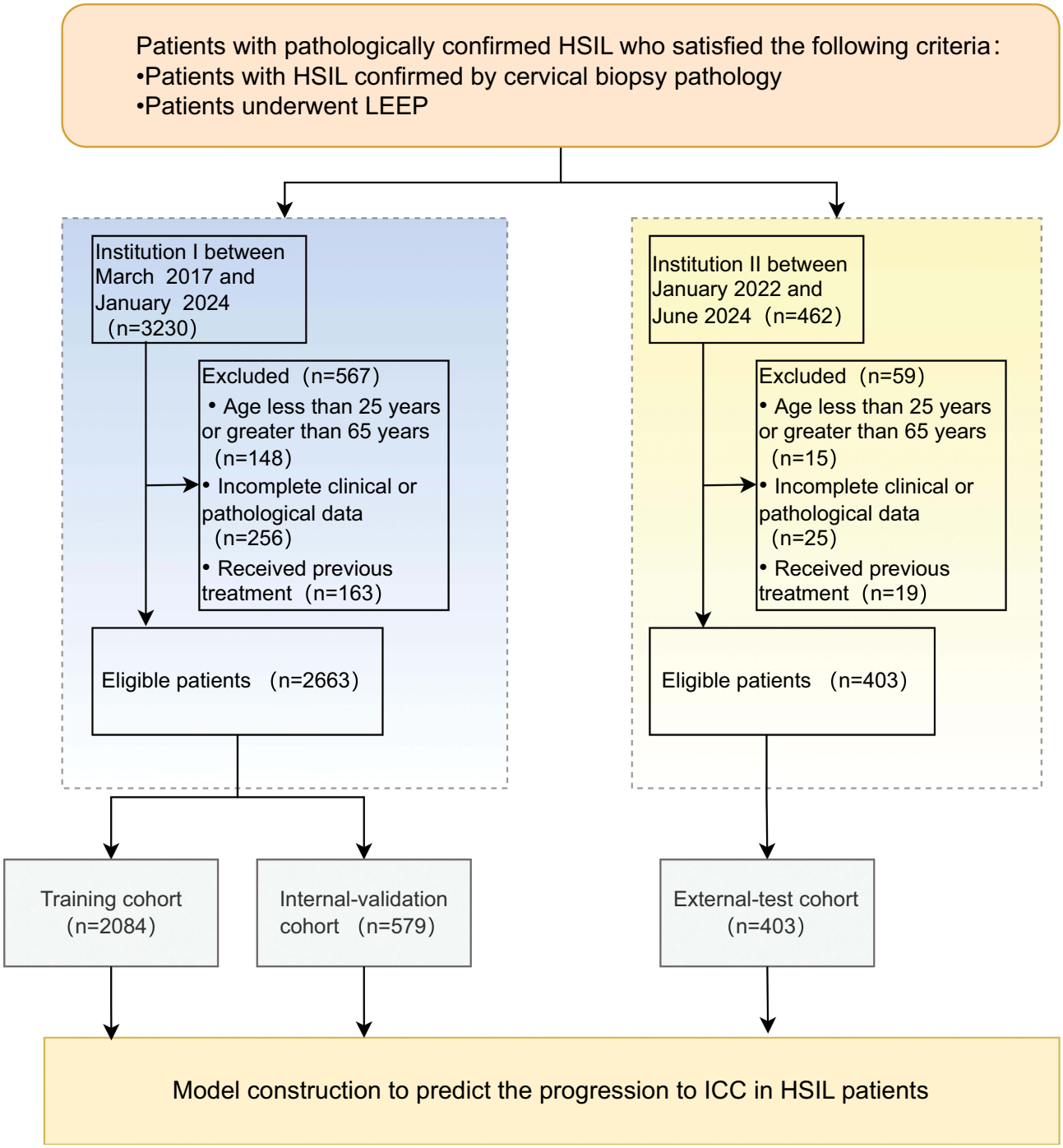

Figure 1 Flowchart of inclusion and exclusion criteria for eligible patients in the study.

https://doi.org/10.2147/RMHP.S536347

endocervical curettage (ECC). HPV status was detected using real-time polymerase chain reaction (PCR). TCT results were classified according to The Bethesda System (TBS). The Transformation Zone (TZ) was categorized into Types I–III based on visibility, following international colposcopic terminology. Nonprogression was defined as follows: the biopsy pathology result was HSIL, and the post-conization pathology result was either HSIL or less (including cervical chronic inflammation, low-grade squamous intraepithelial lesion, or HSIL).[6] Microscopically diagnosed MIC exhibits a stromal invasion depth of less than 5 mm and is characterized by the absence of lymphovascular space invasion.[2]

Clinical Factor Analysis and Modeling

We employed four methods, including Random Forest (RF), Least Absolute Shrinkage and Selection Operator (Lasso), Boruta, and Extreme Gradient Boosting (XGBoost), to screen and select the most predictive variables from the pool of candidates for subsequent modeling analysis. RF is an ensemble learning method combining multiple decision trees to enhance feature selection. It operates by creating diverse trees through bootstrap sampling and random feature subset selection at each node split.[16,17] For clinical feature selection, RF excels by providing variable importance measures, handling high-dimensional data, capturing non-linear relationships, and demonstrating robustness against overfitting.[18] Lasso performs feature selection by adding an L1 penalty term that shrinks coefficients to exactly zero, thereby identifying the most relevant clinical predictors, preventing overfitting, and improving interpretability. Lasso can effectively handle multicollinearity among clinical variables and produces sparse models with enhanced generalizability.[19,20] Boruta is a robust wrapper-based feature selection method that identifies all-relevant predictors by iteratively comparing original features with randomized "shadow" counterparts.[21] It integrates random forest to assess variable importance, retaining attributes consistently outperforming their shuffled versions through binomial testing. Boruta can capture comprehensive feature associations with target variables, which is crucial in clinical studies.[22] XGBoost is an advanced gradient boosting algorithm that builds decision trees sequentially, with each new tree correcting errors from previous ones. It excels in capturing complex non-linear relationships, managing imbalanced clinical data, and delivering high predictive accuracy while preventing overfitting.[23]

To obtain the most robust and clinically significant predictors, we determined the intersection of features selected by the above four methods. This integrative approach minimizes algorithm-specific biases and ensures that the final variables demonstrate strong predictive capability across different machine learning methods.

Model Development and Comparison

We selected 4 widely used ML algorithms, including Logistic Regression (LR), Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM), to comprehensively evaluate our hypothesis. LR models the probability of a binary outcome based on a linear combination of features, offering simplicity and interpretability. XGBoost, an ensemble method, sequentially builds decision trees to correct prior errors, excelling in handling complex datasets with its high accuracy and regularization to prevent overfitting. RF generates multiple decision trees with random data subsets, enhancing model stability and reducing overfitting, making it well-suited for high-dimensional data. SVM identifies the hyperplane that best separates classes, effectively handling high-dimensional and non-linearly separable data using kernel functions.

The efficacy of each model was assessed through comprehensive performance metrics, including receiver operating characteristic (ROC) curve analysis with corresponding area under the curve (AUC) values. Additional evaluation parameters encompassed accuracy, F1 score, and discriminative indices (sensitivity and specificity). Following model comparison, we employed SHapley Additive exPlanations (SHAP) methodology to elucidate the underlying prediction mechanisms and interpret feature importance within the optimal model framework.

Statistics

All statistical tests used in this study were performed using Python software (version 3.11.3) or R software (version 4.2.1). Univariate analysis was conducted using the Chi-square test. A two-sided significance level of $P < 0.05$ was considered statistically significant. For the univariate analyses, we used the autoReg package (version 0.3.3). For the

construction and evaluation of machine learning models, we utilized the Scikit-learn package (version 1.5.2), which provided a robust framework for model development and validation.

Ethics

The requirement for informed consent was waived by the Ethics Committee of Zhangzhou Hospital Affiliated with Fujian Medical University (NO.2024LWB245) and the Qilu Hospital, Shandong University (NO. KYLL-202107-077-1) due to the retrospective nature of the study. All patient data were fully anonymized prior to analysis, ensuring that no identifiable information was used. The study strictly adhered to applicable data protection and privacy regulations, and was conducted in accordance with the principles outlined in the Declaration of Helsinki.

Result

Baseline Characteristics of the Study Cohorts

Table 1 summarizes the baseline clinical characteristics of the 3066 patients included in the study, of whom 299 (9.75%) were diagnosed with MIC. Among these, 9.80% (261/2663) were from Institution I, and 9.42% (38/403) were from

Table 1 Clinical Data Baseline Table of This Study

Variable		Training Cohort			Internal Validation Cohort			External Validation Cohort		
		HSIL (N=1896)	MIC (N=214)	P-value	HSIL (N=506)	MIC (N=47)	P-value	HSIL (N=365)	MIC (N=38)	P-value
Age	<48	1314 (69.3)	102 (47.7)	<0.001	331 (65.4)	23 (48.9)	0.026	294 (80.5)	12 (31.6)	<0.001
	≥48	582 (30.9)	112 (52.3)		175 (34.6)	24 (51.1)		71 (19.5)	26 (68.4)	
TCT≥ASCUS	No	216 (11.4)	15 (7)	=0.054	48 (9.5)	4 (8.5)	0.827	102 (27.9)	9 (18.4)	0.594
	Yes	1680 (88.6)	199 (93)		458 (90.5)	43 (91.5)		263 (72.1)	29 (76.3)	
TCT≥ HSIL	No	1509 (79.6)	114 (53.3)	<0.001	357 (70.6)	26 (55.3)	0.033	315 (86.3)	28 (73.7)	0.043
	Yes	387 (20.4)	100 (46.7)		149 (29.4)	21 (44.7)		50 (13.7)	10 (26.3)	
HPV16/18	No	1049 (55.3)	77 (36)	<0.001	294 (58.1)	21 (44.7)	0.078	168 (58.5)	7 (18.4)	0.002
	Yes	847 (44.7)	137 (64)		212 (41.9)	26 (55.3)		197 (41.5)	31 (81.6)	
Menopause	No	1549 (81.7)	154 (72)	<0.001	382 (75.5)	28 (59.6)	0.019	324 (77.6)	18 (47.4)	<0.001
	Yes	357 (18.3)	90 (42.1)		124 (24.5)	19 (40.4)		41 (22.4)	20 (52.6)	
Gravidity	<3	824 (43.5)	95 (44.4)	=0.794	224 (44.6)	21 (44.7)	0.957	146 (40.0)	15 (39.5)	0.939
	≥3	1072 (56.5)	119 (55.6)		282 (55.7)	26 (55.3)		219 (60.0)	23 (60.5)	
Parity	<3	1614 (85.1)	175 (81.8)	0.197	426 (84.2)	38 (80.9)	0.552	341 (93.4)	35 (92.1)	0.761
	≥3	282 (14.9)	39 (18.2)		80 (15.8)	9 (19.1)		24 (6.6)	3 (7.9)	
TZ III	No	1420 (74.9)	117 (54.7)	<0.001	292 (57.7)	26 (55.3)	0.751	254 (69.6)	9 (23.7)	<0.001
	Yes	476 (25.1)	97 (45.3)		214 (42.3)	21 (44.7)		111 (30.4)	29 (76.3)	
ECC	Negative	1537 (81.1)	91 (42.5)	<0.001	386 (76.3)	17 (36.2)	<0.001	324 (88.8)	20 (52.6)	<0.001
	Positive	359 (18.9)	123 (57.5)		120 (23.7)	20 (63.8)		41 (11.2)	18 (47.4)	
Margin	Negative	1762 (92.9)	124 (57.9)	<0.001	452 (89.3)	21 (44.7)	<0.001	315 (86.3)	25 (65.8)	0.002
	Positive	134 (7.1)	90 (42.1)		54 (10.7%)	26 (55.3)		50 (13.7)	13 (34.2)	

Abbreviations: HSIL, high-grade squamous intraepithelial lesion; MIC, microinvasive cervical cancer; TCT, Thinprep cytologic test; ASCUS, atypical squamous cells of undetermined significance; HPV, human papillomavirus; TZ, Transformation Zone; ECC, endocervical curettage.

Institution II. Patients from Institution I treated before 2022 constituted the training cohort, with MIC prevalence of 11.27% (214/1896). Those from Institution I treated after 2022 formed the internal validation cohort, with MIC prevalence of 9.29% (47/506). Additionally, patients from Institution II were used as an external validation cohort.

Variable Screening

RF algorithm generated 500 decision trees, with each tree randomly selecting four predictive variables at each split. The Mean Decrease Gini (MDG) index was adopted as the indicator for ranking variable importance. Ultimately, the most influential predictive variables associated with the risk of microinvasion were identified and ranked in descending order of importance (Figure 2A). In the Lasso regression model, six predictive variables with non-zero coefficients were identified, and their relative importance rankings are illustrated in Figure 2B. The Boruta algorithm iteratively removed statistically irrelevant features, identifying variables that showed strong or weak correlations with the outcome. After 400 iterations, eight important predictive variables were confirmed (Figure 2C). The XGBoost algorithm provided advantages such as faster computation speed, accurate training results, fewer data constraints, enhanced generalization ability, and improved scalability. This model identified ten significant predictive variables, and their relative importance is illustrated in Figure 2D. The predictive variables identified by RF, Lasso, Boruta, and XGBoost were selected as the common elements for developing the prediction model. Six predictive variables were included: Margin, ECC, TCT≥ HSIL, HPV16/18, TZ III and Age. By using the collinearity diagnosis, the variance inflation factor (VIF) for all the variables were less than 5 (Margin: 1.02; ECC: 1.07; TCT≥ HSIL: 1.06; HPV16/18:1.01; TZ III:1.21; Age: 1.27), indicating no severe collinearity existing in these variables (Figure 3A). We constructed a correlation matrix using Spearman's rank correlation to assess relationships among the six selected variables, as shown in Figure 3B. The analysis revealed

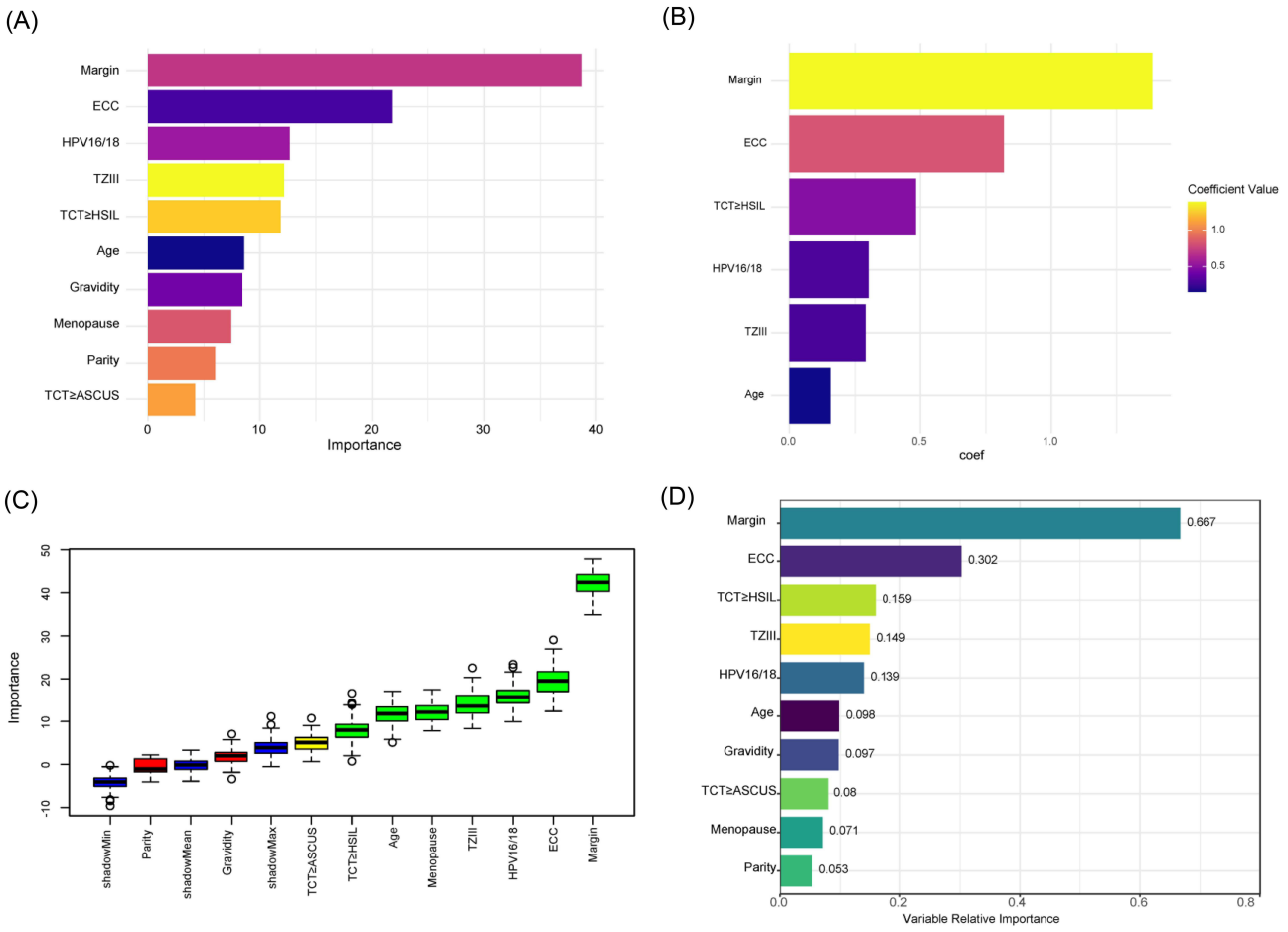

Figure 2 The selected features of different methods. (**A**) RF. (**B**) Lasso. (**C**) Boruta. (**D**) XGBoost.

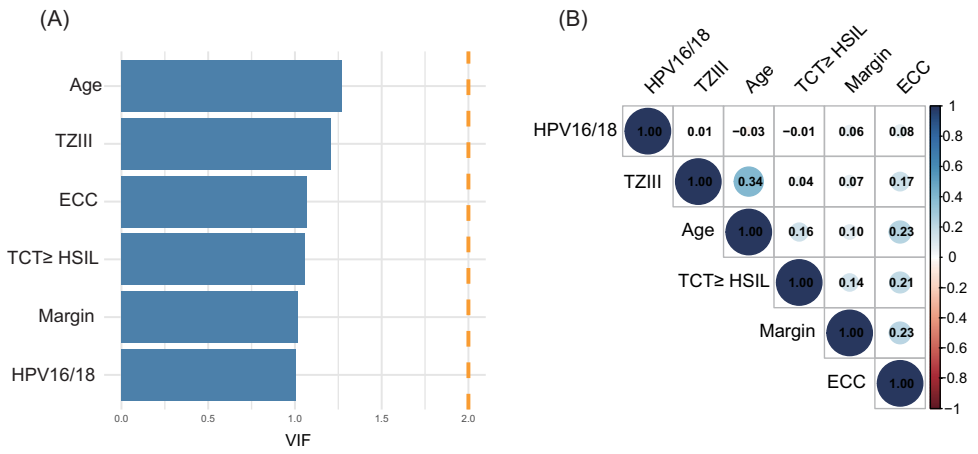

Figure 3 (**A**) VIFs for all the selected variables. (**B**) Correlation matrix.

moderate correlations between Margin and ECC (ρ=0.23), ECC and Age (ρ=0.23), TZ III and Age (ρ=0.34), and TCT≥HSIL and ECC (ρ=0.21). All other pairwise correlations exhibited |ρ|<0.20, indicating low to moderate inter-variable associations.

Prediction Models Development and Validation

To develop robust predictive models for MIC risk in patients with HSIL, we evaluated four machine learning algorithms: LR, SVM, RF, and XGBoost. These models were trained on the training cohort using the six selected predictive variables. To address the class imbalance in the dataset, we implemented class weighting strategies to mitigate bias toward the majority non-MIC class. The decision threshold for all models was set to 0.5, with performance metrics including area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and F1 score reported in Table 2. In the training cohort, all four models exhibited comparable performance, with minor differences observed. The XGBoost algorithm achieved the highest AUC (0.822; 95% CI: 0.793–0.852) and sensitivity (0.778). Meanwhile, the LR model demonstrated superior performance in terms of accuracy (0.777), specificity (0.785), and F1 score (0.383). In the validation cohort, the XGBoost and SVM models both attained identical AUC values (0.827), though their performance differed across other metrics. XGBoost achieved the highest sensitivity (0.851), while the LR model showed balanced performance in accuracy and specificity, both reaching 0.745. Additionally, the RF model

Table 2 Prediction Performance of Different Models

Cohort	Model	AUC (95% CI)	Accuracy±S.E.	Sensitivity±S.E.	Specificity±S.E.	F1 Score±S.E.
Training	XGBoost	0.822 (0.793–0.852)	0.738±0.008	0.778±0.026	0.734±0.009	0.368±0.018
	LR	0.814 (0.784–0.843)	0.777±0.008	0.705±0.029	0.785±0.008	0.383±0.020
	SVM	0.815 (0.786–0.844)	0.753±0.008	0.728±0.028	0.755±0.009	0.366±0.019
	RF	0.812 (0.783–0.841)	0.767±0.008	0.713±0.029	0.773±0.009	0.375±0.020
Internal validation cohort	XGBoost	0.827 (0.763–0.891)	0.691±0.019	0.851±0.051	0.676±0.021	0.319±0.038
	LR	0.817 (0.753–0.880)	0.745±0.018	0.745±0.065	0.745±0.019	0.332±0.043
	SVM	0.827 (0.765–0.889)	0.714±0.019	0.787±0.060	0.708±0.019	0.319±0.040
	RF	0.815 (0.750–0.880)	0.736±0.019	0.787±0.060	0.731±0.019	0.336±0.042
External validation cohort	XGBoost	0.802 (0.730–0.874)	0.734±0.023	0.667±0.078	0.742±0.022	0.327±0.047
	LR	0.801 (0.728–0.873)	0.772±0.022	0.590±0.078	0.791±0.020	0.333±0.049
	SVM	0.771 (0.690–0.852)	0.747±0.023	0.564±0.079	0.766±0.021	0.301±0.048
	RF	0.789 (0.716–0.862)	0.757±0.023	0.564±0.079	0.777±0.021	0.310±0.049

Abbreviations: AUC, area under the curve; XGBoost, eXtreme Gradient Boosting; LR, logistic regression; SVM, support vector machine; RF, random forest; CI, confidence interval; S.E., standard error.

https://doi.org/10.2147/RMHP.S536347

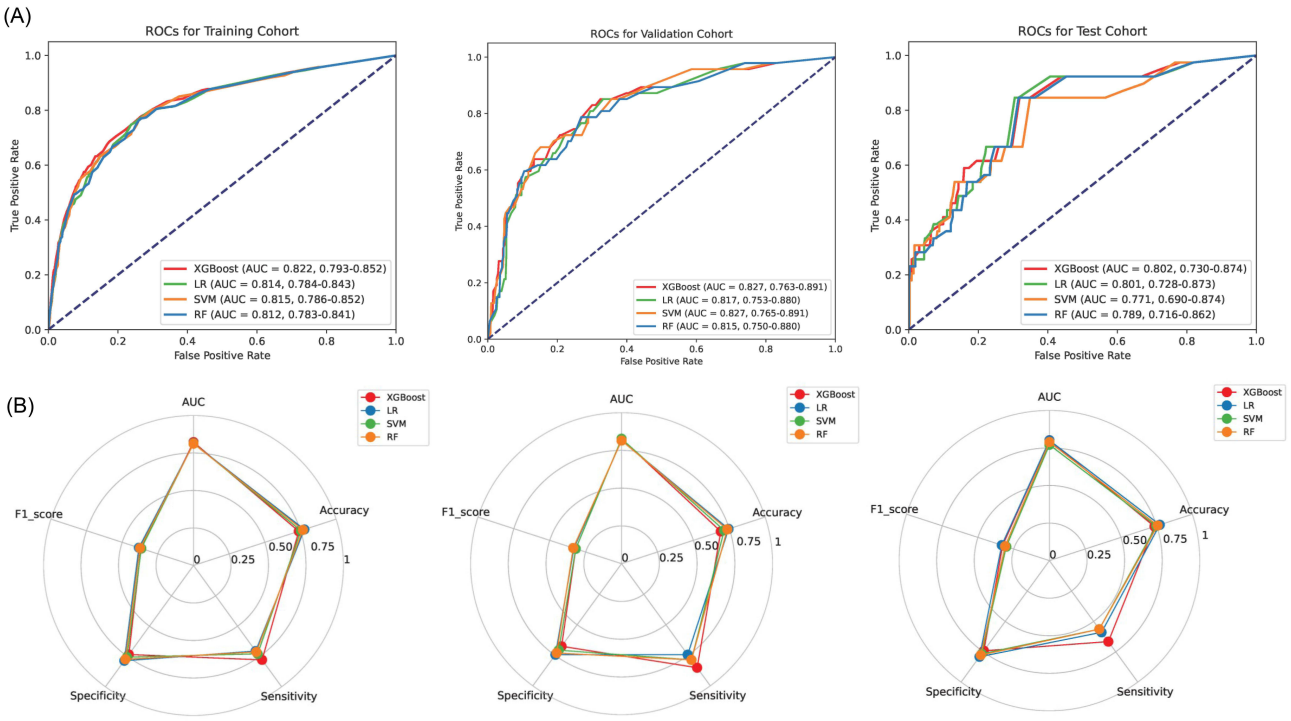

Figure 4 (**A**) Receiver operating characteristic (ROC) curves. Area under the curve (AUC) and 95% CIs of the training, validation, and test cohorts for different machine learning models. (**B**) The radar chart visualization of prediction performance for different models.

exhibited the highest F1 score (0.336) in this cohort. Performance generally declined in the test cohort, particularly regarding sensitivity. The XGBoost and LR models yielded similar AUC values (0.802 and 0.801, respectively). The RF model achieved the highest accuracy (0.757) and specificity (0.777). Overall, despite variations in performance across cohorts and evaluation metrics, the XGBoost algorithm demonstrated consistent performance regarding AUC and sensitivity. Consequently, we selected the XGBoost model as the optimal predictive model, with corresponding AUC values of 0.822, 0.827, and 0.802 across the training, validation, and test cohorts, respectively (Figure 4).

XGBoost Combined Model for SHAP

We calculated the overall and individual Shapley values for the XGBoost model to elucidate its prediction mechanisms and interpret feature importance. The SHAP bar chart (Figure 5A) highlights the six key features—Margin, ECC, TCT≥HSIL, HPV16/18, TZ III, and Age—with average Shapley values of 0.56, 0.55, 0.37, 0.3, 0.16, and 0.09, respectively, underscoring Margin as the strongest contributor. The SHAP beeswarm plot (Figure 5B) illustrates how high values of Margin, ECC, TCT≥HSIL, and HPV16/18 (in red) drive the model toward predicting MIC, consistent with clinical evidence that positive margins and ECC signal residual disease risk. To provide deeper insight into individual predictions, Figure 6 presents force plots for two representative cases: (1) a low-risk case (all variables negative; predicted MIC probability=0.18), where negative ECC and Margin strongly favor non-MIC, and (2) a high-risk case (positive Margin, positive ECC, all other variables negative; predicted MIC probability=0.83), where these positive features dominate the MIC prediction. The SHAP heatmap (Figure 5C) and decision plot (Figure 5D) further elucidate the direction and intensity of each feature's impact across cases and the stepwise contribution to the final predicted probability, respectively.

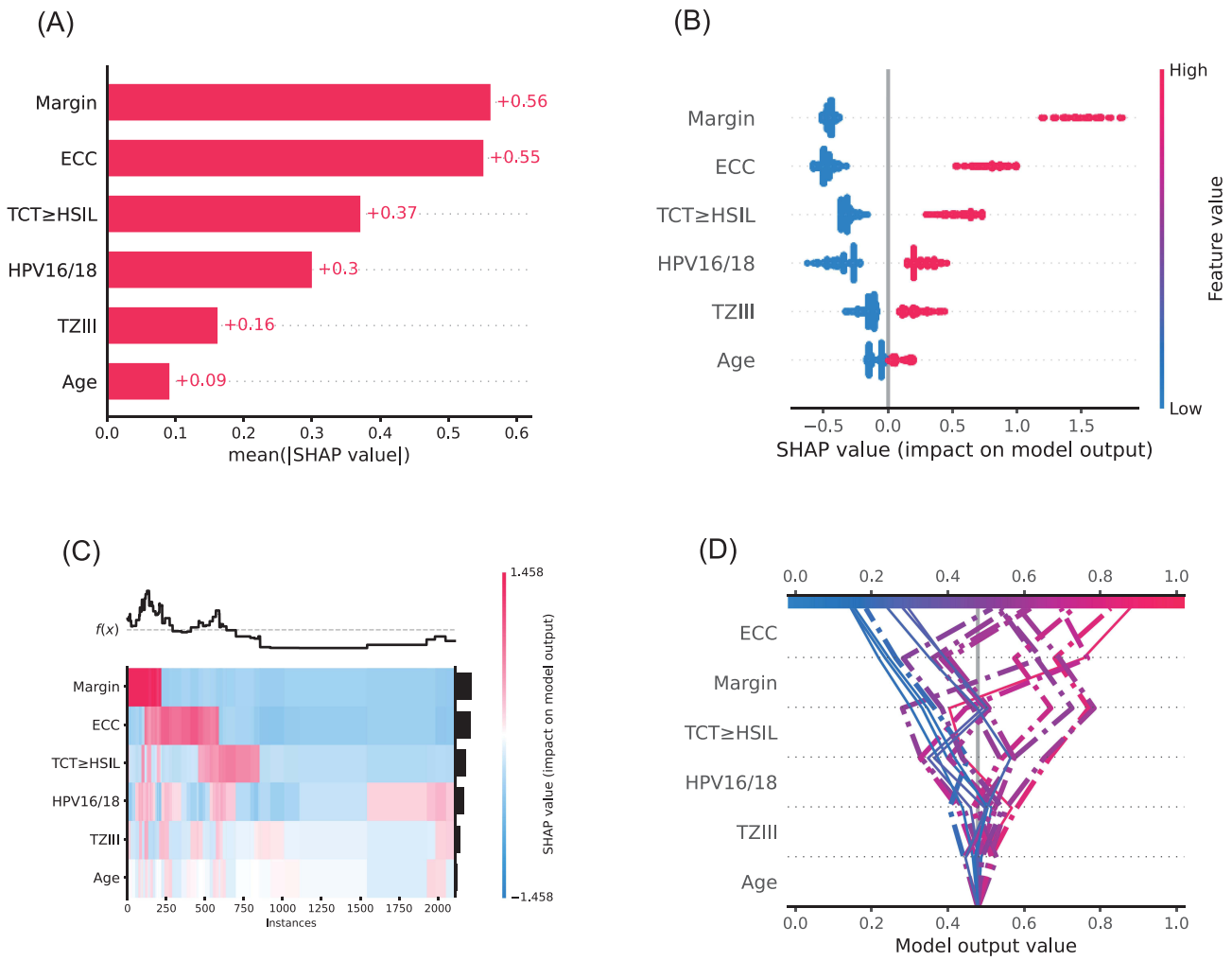

Figure 5 Overall visualization of the model through SHAP. (**A**) The SHAP bar chart shows the weight of the six most important characteristics in the model. (**B**) The SHAP bees-warm plot shows the positive or negative effects of each feature on the prediction probability through red and blue colors. (**C**) The SHAP heatmap plot shows the direction and intensity of influence for each feature of all cases in the model. (**D**) The SHAP decision plot shows the impact process of each significant feature on the final predicted probability.

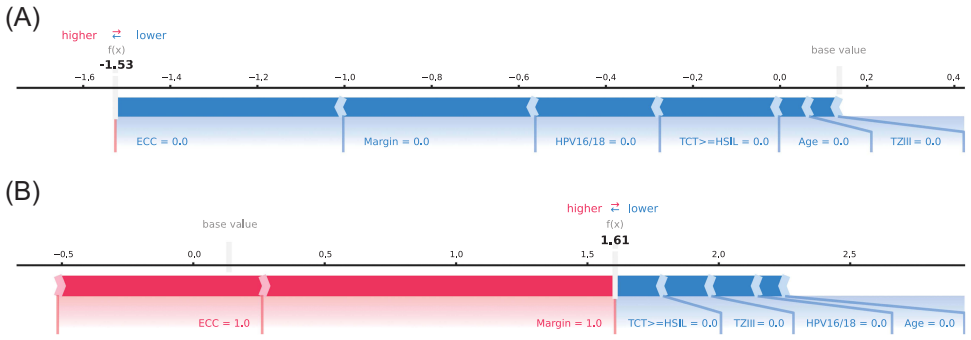

Figure 6 SHAP analysis results for different patient groups. (**A**) True Negative Patient; (**B**) True Positive Patient.

Convenient Application for Clinical Utility

The final predictive model has been integrated into a user-friendly web application designed for clinical settings, as illustrated in Figure 7. Upon entering the actual values for the six predictive features required by the model, the application automatically calculates the patient's risk of progression to invasive cancer. Additionally, the application

https://doi.org/10.2147/RMHP.S536347

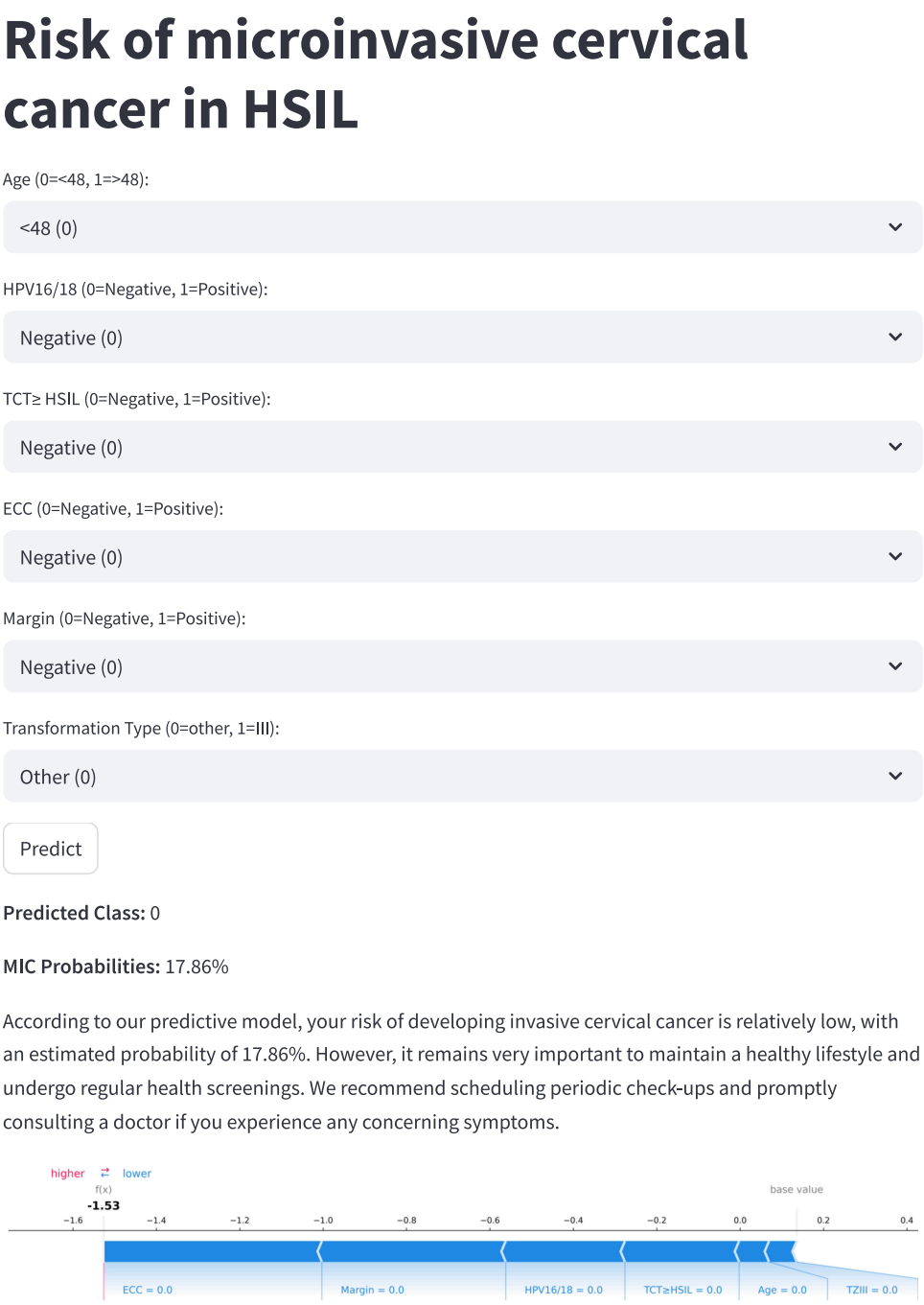

Figure 7 Convenient application for clinical utility. The convenient application of the final XGBoost model with 6 features is available for MIC prediction. When entering actual values of the 6 features, this application automatically displays the predicted probability. Meanwhile, the force plot for individual child indicates the features that contribute to the decision of "MIC": the blue features on the right are the features pushing the prediction towards the "non- MIC" class, while the red features on the left are pushing the prediction towards the "MIC" class.

generates an individualized force plot, visually illustrating the contribution of each feature to the prediction outcome: features displayed in blue support classification as "non-MIC", whereas those shown in red drive the prediction toward "MIC". When the model predicts a high risk, the following message is provided to the patient:

> According to our predictive model, you have a high risk of progression to MIC. Although this result is an estimate based on the model's calculations, it suggests a significant potential risk. I strongly recommend that you consult a gynecological specialist as soon as possible for further evaluation, accurate diagnosis, and timely management or treatment if necessary.

This web application is accessible online at https://microinvasion.streamlit.app/.

Discussion

Cervical high-grade squamous intraepithelial lesions (HSIL) carry a risk of progression to cancer.[24] Currently, predicting whether HSIL will progress to invasive lesions remains challenging, as some patients may already have occult cervical cancer.[6] For young women, it is crucial to balance the potential benefits of treatment with future pregnancy risks, aiming to reduce or minimize the scope of resection surgery to preserve fertility.[25–27] For older individuals unsuitable for cervical conization or hysterectomy-related procedures, assessing cervical cancer risk is essential to avoid missed diagnoses and optimize surgical planning. This study developed a machine learning (ML) model based on multi-center electronic medical record (EMR) data, demonstrating strong discriminatory power and clinical utility in predicting the risk of HSIL progression to MIC. The model has also undergone external validation at another medical institution.

Effective feature selection is critical in developing robust machine learning models for clinical applications, as it ensures the identification of clinically relevant predictors while minimizing noise and overfitting. Our study employed an intersection-based strategy across four feature selection methods to select six robust predictors (Margin, ECC, TCT≥HSIL, HPV16/18, TZ III, Age) for MIC risk prediction. This approach is analogous to the stacked-importance ensemble described by Teo et al[28] for predicting head-and-neck lymphedema, which aggregates feature rankings from multiple base learners to mitigate multicollinearity and overfitting in high-dimensional, low-sample settings. In our study, the intersection method prioritized stability by retaining only features consistently identified across all methods, reducing algorithm-specific biases but potentially excluding subtler predictors compared to weighted stacking, which averages rankings for a more inclusive feature set. For feature importance fusion, our intersection approach ensured high stability, suitable for our imbalanced, multicenter dataset. These findings highlight the value of ensemble feature selection in clinical prediction, and future studies could explore hybrid intersection-stacking approaches to balance stability and inclusivity for enhanced predictive performance.

Preoperative assessment of invasive lesions poses challenges and controversies for clinicians. Previous studies suggest that certain clinical factors may correlate with HSIL progression to invasive cancer. Liu et al developed a clinical prediction model for pathological upgrading to invasive carcinoma after cervical conization. Independent risk factors included age, contact bleeding symptoms, HPV16/18 infection, HSIL cytology, pathological grading of HSIL in cervical biopsy, suspected stromal invasion on pathology, and HSIL in endocervical curettage (ECC).[6] This model provides a robust, clinically valuable tool for predicting HSIL progression risk, guiding preoperative risk assessment and surgical strategy optimization. Liu et al further confirmed through multivariate logistic regression that age, contact bleeding, HPV16/18 infection, HSIL cytology, pathological grading, and suspected stromal invasion are high-risk factors for pathological upgrading post-conization, suggesting disease severity and persistent HPV infection may correlate with progression. These findings align closely with our results. Margin status, ECC, TCT≥HSIL, HPV16/18 infection, transformation zone involvement, and age were identified as predictive variables for progression to MIC after LEEP.

Among the predictive variables we identified, the resection completeness of HSIL is defined by margin status, where positive margins significantly correlate with residual and recurrent disease.[29] With advancing age, the squamocolumnar junction migrates deeper into the endocervical canal, becoming less accessible to colposcopic visualization. In cases where HSIL coexist with MIC, the malignant transformation typically originates at the superior margin of the HSIL. Clinically significant is the finding that 41.6–57.4% of cervical carcinomas either involve a non-visualized squamocolumnar junction or exhibit endocervical extension of the lesion. These observations collectively suggest that advanced age, TZIII, and ECC positivity constitute independent risk factors for neoplastic progression from HSIL to MIC.[30]

We employed SHAP to visualize feature importance and model behavior. The SHAP bar plot identified the top 6 predictive features: Margin status, ECC, TCT≥HSIL, HPV16/18 infection, transformation zone involvement, and age. The SHAP summary plot demonstrated that positive SHAP values (red) indicate increased MIC risk, negative SHAP values (blue) suggest protective effects. The Margin status showed the strongest positive correlation, followed by ECC. The analysis offers clinical insights into the model's decision-making patterns, with Margin status and positive ECC emerging as the most significant predictive features for MIC after LEEP risk. Furthermore, the SHAP summary heatmap illustrates both the directional impact (positive or negative) and the relative importance of each feature across all observed instances in the dataset.

In this study, we integrated and expanded upon key clinical features identified in previous research. Compared to existing models, we employed four machine learning algorithms—random forest (RF), Lasso, Boruta, and XGBoost—to screen six influential variables from the dataset. The XGBoost-derived model demonstrated superior performance in AUC of 0.822 (95% CI: 0.793–0.852) and sensitivity (0.778), with robust validation across internal and external cohorts. This novel model achieves higher predictive accuracy through rigorous variable selection, development, validation, interpretation, and web-based deployment. Its accessibility and practicality offer significant clinical value, enabling rapid decision-making. By assessing real-time cervical cancer risk preoperatively, the model guides tailored surgical approaches—such as deferring or opting for ablation in low-risk cases to preserve fertility, versus designing adequate resection margins in high-risk cases to avoid repeat surgeries.[31] These strategies may reduce intraoperative complications, enhance surgical success, and optimize postoperative care.

It is widely acknowledged that ML models offer superior predictive performance compared to traditional linear models.[32] This advantage enables the construction of a relatively robust model from complex data.[33] The six predictors in this study, such as margin status and ECC, exemplify the handling of multidimensional heterogeneity. Manual curation of EMR data ensured accuracy, while ML-driven predictive modeling supports personalized medicine and care quality improvement.[34] Visualization strategies and a user-friendly web application were developed to demystify the "black-box" nature of ML, fostering trust in clinical applications.

Our study has several limitations. First, the retrospective design may introduce selection bias. Second, despite multicenter data, the low incidence of post-LEEP progression limited sample sizes, potentially affecting external validation performance. Third, missing key data (eg, colposcopic imaging) may have influenced results. Fourth, prospective validation of the ML model is pending, representing a future focus for our team.

Conclusion

We developed and validated a non-invasive and robust machine learning-based ensemble model using clinical information features to identify the risk of progression to MIC in HSIL patients after LEEP surgery. SHAP provides a bridge for personalized prediction, which may help offer personalized treatment and clinical decision-making support for HSIL patients preparing to undergo LEEP procedures. Future research should focus on developing standardized risk assessment tools and validating their clinical utility in diverse patient populations. This will ultimately lead to improved patient outcomes and more efficient use of healthcare resources in the management of early-stage cervical malignancies.

Data Sharing Statement

All the data underlying this article are presented in the article and available from the corresponding author upon reasonable request.

Ethics

This study has been reviewed and approved by the Ethics Committee of Zhangzhou Hospital Affiliated with Fujian Medical University (NO.2024LWB245), the Qilu Hospital, Shandong University (NO. KYLL-202107-077-1). Given the retrospective nature of this study, the requirement for obtaining informed consent was waived. This study was performed in line with the principles of the Declaration of Helsinki.

Funding

This study was supported by the Start-Up Fund of Fujian Medical University (2020QH1279), Young Talent of Lifting engineering for Science and Technology in Shandong, China (SDAST2025QTB021), the Jinan City "20 New Universities" independent innovation group (2021GXRC027), Natural Science Foundation of Shandong Province of China (ZR2023QH141), Natural Science Foundation of China (82303750), Medical and Health Science and Technology Development Plan of Shandong Province (202105010417).

Disclosure

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhang T, Zhuang L, Muaibati M, et al. Identification of cervical cancer stem cells using single-cell transcriptomes of normal cervix, cervical premalignant lesions, and cervical cancer. *eBioMedicine*. 2023;92:104612. doi:10.1016/j.ebiom.2023.104612
2. Fanghui Z, Youlin Q. Cervical cancer prevention in China: a key to cancer control. *Lancet*. 2019;393(10175). doi:10.1016/s0140-6736(18)32849-6
3. Wang H, Liu C, Jin K, Li X, Zheng J, Wang D. Research advances in signaling pathways related to the malignant progression of HSIL to invasive cervical cancer: a review. *Biomed Pharmacother*. 2024;180:117483. doi:10.1016/j.biopha.2024.117483
4. Peto J, Gilham C, Fletcher O, Matthews FE. The cervical cancer epidemic that screening has prevented in the UK. *Lancet*. 2004;364 (9430):249–256. doi:10.1016/S0140-6736(04)16674-9
5. Reich O, Regauer S, Gutierrez AL, Kashofer K. Copy number profiling implicates thin high-grade squamous intraepithelial lesions as a true precursor of cervical human papillomavirus-induced squamous cell cancer. *Lab Invest*. 2024;104(9):102108. doi:10.1016/j.labinv.2024.102108
6. Liu Q, Yang J, Cheng H, Shu C, Tang Y, Zhao J. A clinical prediction model for pathologic upgrade to invasive carcinoma following conization of cervical high-grade squamous intraepithelial lesions. *Cancer Med*. 2024;14(1). doi:10.1002/cam4.70540
7. Zhang M, Lin X, Zheng Z, Chen Y, Ren Y, Zhang X. Artificial intelligence models derived from 2D transperineal ultrasound images in the clinical diagnosis of stress urinary incontinence. *Int Urogynecol J*. 2022;33(5):1179–1185. doi:10.1007/s00192-021-04859-y
8. Chou B, Prasad Venkatesulu B, Coleman RL, Harkenrider M, Small W. Management of stage I and II cervical cancer: a review. *Int J Gynecol Cancer*. 2022;32(3):216–224. doi:10.1136/ijgc-2021-002527
9. Zhang L, Tian P, Li B, et al. Risk-stratified management of cervical high-grade squamous intraepithelial lesion based on machine learning. *J Med Virol*. 2024;96(10):e70016. doi:10.1002/jmv.70016
10. Dou Y, Zhang X, Li Y, Wang F, Xie X, Wang X. Triage for management of cervical high-grade squamous intraepithelial lesion patients with positive margin by conization: a retrospective analysis. *Front Med*. 2017;11(2):223–228. doi:10.1007/s11684-017-0517-8
11. Shi Y, Zhang G, Ma C, et al. Machine learning algorithms to predict intraoperative hemorrhage in surgical patients: a modeling study of real-world data in Shanghai, China. *BMC Med Inf Decis Making*. 2023;23(1). doi:10.1186/s12911-023-02253-w
12. Hu J, Xu J, Li M, et al. Identification and validation of an explainable prediction model of acute kidney injury with prognostic implications in critically ill children: a prospective multicenter cohort study. *eClinicalMedicine*. 2024;68:102409. doi:10.1016/j.eclinm.2023.102409
13. Chen X, Zhang H, Guo D, et al. Risk of intraoperative hemorrhage during cesarean scar ectopic pregnancy surgery: development and validation of an interpretable machine learning prediction model. *eClinicalMedicine*. 2024;78:102969. doi:10.1016/j.eclinm.2024.102969
14. Azodi CB, Tang J, Shiu S-H. Opening the black box: interpretable machine learning for geneticists. *Trends Genet*. 2020;36(6):442–455. doi:10.1016/j.tig.2020.03.005
15. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:1.
16. Farhadian M, Torkaman S, Mojarad F. Random forest algorithm to identify factors associated with sports-related dental injuries in 6 to 13-year-old athlete children in Hamadan, Iran-2018 -a cross-sectional study. *BMC Sports Sci Med Rehabil*. 2020;12(1):69. doi:10.1186/s13102-020-00217-5
17. Shi G, Liu G, Gao Q, et al. A random forest algorithm-based prediction model for moderate to severe acute postoperative pain after orthopedic surgery under general anesthesia. *BMC Anesthesiol*. 2023;23(1). doi:10.1186/s12871-023-02328-1
18. Nachouki M, Mohamed EA, Mehdi R, Abou Naaj M. Student course grade prediction using the random forest algorithm: analysis of predictors' importance. *Trends Neurosci Educ*. 2023;33:100214. doi:10.1016/j.tine.2023.100214
19. Wang J, Xu Y, Liu L, et al. Comparison of LASSO and random forest models for predicting the risk of premature coronary artery disease. *BMC Med Inf Decis Making*. 2023;23(1). doi:10.1186/s12911-023-02407-w
20. Kang J, Choi YJ, Kim I-K, et al. LASSO-based machine learning algorithm for prediction of lymph node metastasis in T1 colorectal cancer. *Cancer Res Treat*. 2021;53(3):773–783. doi:10.4143/crt.2020.974
21. Zhou H, Xin Y, Li S. A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC Bioinf*. 2023;24(1). doi:10.1186/s12859-023-05300-5
22. Sun Y, Zhang Q, Yang Q, Yao M, Xu F, Chen W. Screening of gene expression markers for corona virus disease 2019 through Boruta_MCFS feature selection. *Front Public Health*. 2022;10. doi:10.3389/fpubh.2022.901602
23. Moore A, Bell M. XGBoost, A novel explainable AI technique, in the prediction of myocardial infarction: a UK Biobank Cohort Study. *Clin Med Insights Cardiol*. 2022;16. doi:10.1177/11795468221133611
24. McCredie MRE, Sharples KJ, Paul C, et al. Natural history of cervical neoplasia and risk of invasive cancer in women with cervical intraepithelial neoplasia 3: a retrospective cohort study. *Lancet Oncol*. 2008;9(5):425–434. doi:10.1016/s1470-2045(08)70103-7
25. Klaritsch P, Reich O, Giuliani A, Tamussino K, Haas J, Winter R. Delivery outcome after cold-knife conization of the uterine cervix. *Gynecol Oncol*. 2006;103(2):604–607. doi:10.1016/j.ygyno.2006.04.003
26. Kyrgiou M, Athanasiou A, Paraskevaidi M, et al. Adverse obstetric outcomes after local treatment for cervical preinvas ive and early invasive disease according to cone depth: systematic rev iew and meta-analysis. *BMJ*. 2016;354:i3633. doi:10.1136/bmj.i3633
27. Zhu M, Yu M, Chen Z, Zhao W. Construction and evaluation of a clinical prediction scoring system fo r positive cervical margins under colposcopy. *Front Med*. 2022;9:807849. doi:10.3389/fmed.2022.807849
28. Teo PT, Rogacki K, Gopalakrishnan M, et al. Determining risk and predictors of head and neck cancer treatment-related lymphedema: a clinicopathologic and dosimetric data mining approach using interpretable machine learning and ensemble feature selection. *Clin Transl Radiat Oncol*. 2024;46:100747. doi:10.1016/j.ctro.2024.100747
29. Arbyn M, Redman CWE, Verdoodt F, et al. Incomplete excision of cervical precancer as a predictor of treatment failure: a systematic review and meta-analysis. *Lancet Oncol*. 2017;18(12):1665–1679. doi:10.1016/s1470-2045(17)30700-3
30. Liu Q, Yang J, Cheng H, Shu C, Tang Y, Zhao J. A clinical prediction model for pathologic upgrade to invasive carcinoma following conization of cervical high-grade squamous intraepithelial lesions. *Cancer Med*. 2025;14(1):e70540. doi:10.1002/cam4.70540

31. Abu-Rustum NR, Yashar CM, Bean S, et al. NCCN guidelines insights: cervical cancer, version 1.2020. *J Natl Compr Canc Netw*. 2020;18 (6):660–666. doi:10.6004/jnccn.2020.0027
32. Ota R, Yamashita F. Application of machine learning techniques to the analysis and predict ion of drug pharmacokinetics. *J Control Release*. 2022;352:961–969. doi:10.1016/j.jconrel.2022.11.014
33. Zhou S-N, Jv D-W, Meng X-F, et al. Feasibility of machine learning-based modeling and prediction using mu ltiple centers data to assess intrahepatic cholangiocarcinoma outcomes. *Ann Med*. 2023;55(1):215–223. doi:10.1080/07853890.2022.2160008
34. Dubovitskaya A, Baig F, Xu Z, et al. ACTION-EHR: patient-centric blockchain-based electronic health record data management for cancer care. *J Med Internet Res*. 2020;22(8):e13598. doi:10.2196/13598