Research Square

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

Risk factors analysis and prediction model establishment of lymph node metastasis in Rectal Cancer based on peripheral blood cell comprehensive score

Jinkun Xie
Lianjiang County General Hospital
Baitong Che
Jinzhou Central Hospital
Wei Zhang
The First Affiliated Hospital of Jinzhou Medical University
Chongshi Gao
Jinzhou Medical University
Zhiyong Yang
Jinzhou Medical University
Weibin Zhang
webean_zhang@163.com

Jinzhou Medical University

Research Article

Keywords: Rectal cancer, Peripheral blood cell score, Lymph node metastasis, Retrospective analysis, Prediction model

Posted Date: September 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10376688/v1

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

Additional Declarations: No competing interests reported.

Risk factors analysis and prediction model establishment of lymph node metastasis in Rectal Cancer based on peripheral blood cell comprehensive score

Jinkun Xie[1], Baitong Che[2], Wei Zhang[3], Chongshi Gao[1], Zhiyong Yang[1], Weibin Zhang[2]*

1 Lianjiang County General Hospital, Fuzhou, Fujian, 350000, China.

2 Jinzhou Central Hospital, Jinzhou, Liaoning, 121000, China.

3 The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, Liaoning, 121001, China.

*** Correspondence:** Weibin Zhang; Email: webean_zhang@163.com

Keywords: Rectal cancer; Peripheral blood cell score; Lymph node metastasis; Retrospective analysis; Prediction model

Abstract

Background: Lymph node metastasis (LNM) is a critical determinant of therapeutic strategies for patients with rectal cancer (RC). This study aimed to identify independent risk factors for LNM and develop machine learning-based predictive models incorporating the comprehensive peripheral blood cell score (PBCS) and routine clinical parameters. Methods: A total of 500 patients who received curative resection for RC were retrospectively enrolled. The PBCS was established, and multivariate regression was performed to screen independent predictive factors. Eight machine learning algorithms were adopted to construct predictive models with multi-dimensional performance validation. After comprehensive assessment of model

efficacy, the optimal algorithm was selected to develop a simplified clinical scoring system and an online predictive calculator. Results: Age, maximum tumor diameter, PBCS, preoperative carcinoembryonic antigen (CEA), and pretreatment cT stage were verified as independent risk factors for RC-related LNM, among which the XGBoost (XGB) algorithm yielded the optimal predictive performance. Conclusion: The finalized XGB-based model enables reliable preoperative risk stratification of LNM in RC patients, and the auxiliary predictive tools provide an objective reference for individualized clinical decision-making

1 Introduction

Colorectal cancer (CRC) represents one of the most prevalent gastrointestinal malignancies worldwide, with a steadily increasing global incidence. According to the World Health Organization (WHO), annual CRC-related mortality exceeds 900,000 globally (1). In East Asia, where this disease is highly prevalent, the lifetime risk of colorectal cancer is 2% (2).While rectal cancer (RC) accounts for a significant proportion of colorectal cancer cases, early screening and diagnosis are crucial for improving patient prognosis. In Japan, standardized early screening programs have lifted the 5-year overall survival of RC patients above 60%. Nevertheless, advanced RC accompanied by distant metastasis substantially complicates clinical management, with the corresponding survival rate dropping below 15% (3). CRC is a highly aggressive cancer with strong invasive and metastatic capabilities, and lymph node metastasis (LNM) is its main route of metastasis. LNM is the pathological process by which tumor cells spread through the lymphatic system, playing a crucial role in

tumor development and dissemination, and having a significant impact on disease treatment and prognosis. Accurately predicting which high-risk patients with rectal cancer are likely to develop lymph node metastasis can improve patients' quality of life and reduce unnecessary waste of medical resources.

In the diagnosis, pre-treatment assessment, efficacy monitoring and follow-up of colorectal cancer, peripheral blood tumor marker detection has important clinical reference value. Commonly used indicators include carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA19-9) (4). For special metastasis types such as liver metastasis, peritoneal metastasis and ovarian metastasis, markers such as alpha fetoprotein (AFP) and carbohydrate antigen 125 (CA125) also have important auxiliary diagnostic significance (5, 6). In addition to tumor markers in peripheral blood, markers of inflammatory response also play a particularly important role in the occurrence and development of tumors. Local and systemic inflammatory reactions can remodel the tumor immune microenvironment, further facilitating malignant cell proliferation, invasion and distant metastasis. Such inflammatory biomarkers mainly consist of neutrophils, lymphocytes, platelets and serum albumin (7). Research confirms that platelets can stimulate tumor cell growth by aggregating and degranulating in tumor microvessels, while tumor-related inflammatory mediators can also reverse-induce platelet elevation (8). Lymphocytes are also important components of antitumor immune responses, capable of precisely identifying and killing tumor cells, or activating the body's antitumor immune response by secreting various cytokines (9). Neutrophils, as primary response cells to inflammation,

infection, and tissue damage, can be recruited and enriched by the tumor microenvironment (TME) through mediators such as cytokines, chemokines, lipids, and growth factors, further promoting tumor cell proliferation and distant metastasis (10). Systemic inflammatory response indicators are an emerging category of serum biomarkers with significant prognostic potential in various hematologic tumors and solid tumors (11, 12). The emerging serum biomarkers included in this study are: neutrophil-lymphocyte ratio (NLR), platelet-lymphocyte ratio (PLR), platelet-albumin ratio (PAR), and hemoglobin-platelet ratio (HPR), with previous studies demonstrating that all of these indicators are closely associated with tumor progression, metastasis, and other adverse prognoses (13-15). Accumulating evidence has confirmed the prognostic utility of these hematological markers, yet large-scale clinical validation is still warranted to refine the RC prognostic scoring system and identify more specific predictive biomarkers. Based on this, this study established a peripheral blood cell comprehensive score (PBCS) (16)by integrating the aforementioned peripheral blood cell ratios, and combined patient general data, imaging data, and serum tumor markers for model training and validation, ultimately selecting the predictive model with the optimal diagnostic efficacy.

The interdisciplinary integration between artificial intelligence and clinical medicine has advanced rapidly in recent decades, and machine learning (ML) has become an indispensable analytical tool in this field. Machine learning (Machine Learning, ML), as an emerging and popular artificial intelligence technology, has been widely applied in medical and health data analysis and clinical decision support scenarios due to its

ability to accurately predict events and outcomes (17). ML can continuously optimize models by autonomously learning data features, ultimately outputting analytical models with predictive and decision-making functions. Compared to traditional statistical methods, machine learning algorithms can analyze data associations more dimensionally, adapting to the in-depth mining of complex clinical data. With the iterative upgrades of computer computing power and storage capacity, machine learning can identify important connections in medical data, promoting personalized treatment recommendations, efficient healthcare services, and cost reduction. Currently, machine learning has been widely applied in cancer metastasis assessment and early prediction research, achieving certain research results.

In summary, this study is based on the constructed comprehensive scoring system for peripheral blood cells to systematically analyze the risk factors for LNM in rectal cancer patients, identify independent risk factors for LNM, and further train eight different machine learning models. The optimal model was then selected as the final predictive model, aiming to provide clinicians with an objective basis for assessing the risk of lymph node metastasis in rectal cancer patients, assist in the development of precision and personalized diagnostic and treatment strategies, and offer scientific references for clinical screening and treatment decision-making for rectal cancer patients.

2 Materials and methods

2.1 Study population and clinical data

This retrospective observational study enrolled patients who underwent curative

radical resection for RC at Jinzhou Central Hospital between January 2019 and December 2024. According to the pre-established inclusion and exclusion criteria, a total of 500 patients were enrolled in the study, and all patients successfully completed the radical rectal cancer surgery. Based on the presence or absence of lymph node metastasis identified in postoperative pathological examination results, patients were divided into a lymph node metastasis-free group (176 cases) and a lymph node metastasis group (324 cases). All research procedures involving human participants in this study comply with the ethical guidelines of the Declaration of Helsinki (2013 revision) and have been reviewed and approved by the medical research ethics committee of our hospital. The ethical approval number is: Hospital Ethics Review No. 2025-LW-034.

During the study implementation, cases with missing or abnormal data were excluded to ensure data quality. Additionally, strict inclusion and exclusion procedures were followed to ensure the completeness and reliability of the enrolled cases' information, providing high-quality data support for constructing a predictive model for assessing the risk of lymph node metastasis in rectal cancer (RC) patients.

Inclusion criteria: (1) Complete clinical data of patients; (2) No prior radiotherapy, chemotherapy, targeted therapy, or any other form of antitumor treatment before surgery; (3) No prior platelet-boosting therapy or other treatments that may affect test indicators before surgery; (4) Postoperative pathological confirmation of no tumor residue at the surgical margins; (5) Preoperative colonoscopy biopsy suggesting RC and the first systemic radical rectal cancer surgery, with postoperative pathological

diagnosis confirming rectal cancer and successful completion of radical rectal cancer surgery; (6) Imaging assessment showing no distant metastasis and no concurrent malignant tumors at other sites; (7) Age ≥ 18 years, excluding minors.

Exclusion criteria: (1) Incomplete clinical data; (2) Presence of multiple primary cancers; (3) Prior radiotherapy, chemotherapy, or other neoadjuvant treatments before surgery; (4) Preoperative presence of infection, hematological diseases, or concurrent severe organ diseases (e.g., heart or lungs) and immunological diseases; (5) Tumors with distant metastasis; (6) History of other malignant tumors, familial adenomatous polyposis, or inflammatory bowel disease.

2.2 Construction of the Peripheral Blood Cell Score System(PBCS)

With lymph node metastasis as the endpoint, ROC curve analysis was performed on four indicators: PLR, NLR, PAR, and HPR. The optimal cutoff values were determined using the Youden index (36): PLR was 178.025, NLR was 3.3975, PAR was 42.565, and HPR was 0.3815. The areas under the curve (AUC) for each indicator were 0.796, 0.797, 0.707, and 0.519, respectively. A single-indicator binary scoring method on a 0–4 point scale was adopted, where "high values" for PLR, NLR, and PAR indicated an increased risk of lymph node metastasis, and "low values" for HPR indicated an increased risk. The specific rules were as follows: if a patient's PLR, NLR, and PAR values were ≥ the corresponding optimal cutoff values, each was scored as 1; if < the corresponding optimal cutoff values, each was scored as 0; if a patient's HPR value was ≤ the corresponding optimal cutoff value, it was scored as 1; if > the corresponding optimal cutoff value, it was scored as 0. The binary

155 classification results of each indicator were summed equally, with the calculation
156 formula being: PBCS = PLR score + NLR score + PAR score + HPR score.

157 **2.3 Research methods**

158 This study collected and organized clinical data from patients including: gender, age,
159 history of diabetes, history of smoking, family history, and history of hypertension;
160 Preoperative hematological indicators include :P LR, NLR, PAR, HPR, FIB, and Hb;
161 Preoperative tumor markers include: CEA, CA19-9, CA72-4, CA125; Preoperative
162 and intraoperative pathological information includes: tumor cT staging, degree of
163 histopathological differentiation, general tumor type, vascular invasion, nerve
164 invasion, microsatellite instability, mismatch repair proteins MSH6, MSH2, MLH1,
165 and PMS2. Based on postoperative pathological reports, whether lymph node
166 metastasis occurred was used as the outcome variable for subsequent statistical
167 analysis and model construction.

168 **2.4 Observation Indicators**

169 General clinical data: (1) Gender: Male, female. (2) Age. (3) History of diabetes:
170 According to the American Diabetes Association (ADA) guidelines, the diagnostic
171 criteria are: 1. Fasting blood glucose ≥ 7.0 mmol/L, measured after at least 8 hours of
172 indigestion. 2. Random blood glucose ≥ 11.1 mmol/L, can be measured at any time
173 and is associated with symptoms such as polyuria, thirst, or weight loss. 3. Oral
174 Glucose Tolerance Test (OGTT): Venous blood is collected 2 hours after a 75 g
175 glucose load. If blood glucose ≥ 11.1 mmol/L (200 mg/dL), diabetes diagnosis is
176 confirmed. 4. HbA1c hemoglobin ≥ 6.5% (48 mmol/mol), reflecting the average

177 blood glucose level over the past 2-3 months. (4) History of smoking: According to
178 WHO, patients with more than or cumulative months of more than 6 months are
179 defined as smokers. (5) History of hypertension: This study referred to the diagnostic
180 criteria of the "Chinese Hypertension Prevention and Treatment Guidelines,"
181 classifying those with systolic blood pressure ≥140 mmHg and/or diastolic blood
182 pressure ≥ 90 mmHg as the hypertension group.
183 Preoperative pathological information: cT staging: The extent of the tumor assessed
184 by non-surgical methods before treatment. Preoperative assessment of tumor invasion
185 depth is made by ultrasound endoscopy, high-resolution MRI, enhanced CT,
186 colonoscopy, biopsy, etc., with at least two senior physicians independently assessed.
187 If ambiguous, a decision is made through multidisciplinary consultation. To reduce
188 errors, tumor cT staging follows the American Joint Committee on Cancer (AJCC)
189 8th edition guidelines: T1 stage involves tumor invasion of the submucosa, T2 stage
190 invades the muscularis propria, stage T3 penetrates the muscularis propria, and stage
191 T4 penetrates the visceral peritoneum or involves adjacent organs. The determination
192 process was independently evaluated by two senior physicians using endoscopic
193 ultrasound, high-resolution MRI, and enhanced CT. If there were disagreements, the
194 decision was determined through multidisciplinary consultation (MDT); In this study,
195 T1+T2 stage was grouped into the shallow infiltration group, and T3+T4 stage into
196 the deep infiltration group. The degree of histopathological differentiation, general
197 tumor type, vascular invasion, nerve invasion, microsatellite instability, mismatch
198 repair proteins MSH6, MSH2, MLH1, and PMS2 were all identified by preoperative

endoscopic biopsy or cytology puncture pathology. General tumor classification: According to the WHO classification standards for digestive tract tumors, tumors are classified as raised, ulcerated, and infiltrated. Among the cases included in this study, only raised and ulcer types were included, with no infiltrative cases collected. (Corresponding code assignment: raised type = 1, ulcer type = 0).

Preoperative hematological indicators: All hematological indicators are taken from the patient's preoperative peripheral blood test results. The patient has not received platelet-increasing drugs or other interventional treatments that may affect test results before surgery, ensuring the accuracy and reliability of the indicators.

2.5 Statistical Analysis

This study's measurement data underwent normality tests, with data conforming to normal distribution represented as $x \pm s$, and intergroup comparisons using independent sample t-tests; data not conforming to normal distribution were represented as median (interquartile range), with intergroup comparisons using Mann-Whitney U tests. Categorical data were presented as number (percentage), with intergroup comparisons using χ^2 tests; when theoretical frequency < 5 or a group within the intergroup had 100% cases, Fisher's exact test was used instead, and the results should be interpreted with caution, as noted in table annotations. All statistical analyses were performed using two-sided tests, and $P < 0.05$ was defined as the threshold for statistical significance. A total of eight classical machine learning classification algorithms were selected to construct predictive models, including Extreme Gradient Boosting (XGB), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Multilayer

Perceptron (MLP), Logistic Regression (LR), Support Vector Machine (SVM), and LightGBM. The aforementioned machine learning models were trained and constructed using Rstudio and R4.5.0 software. During model building, 10-fold cross-validation was used to optimize training set model parameters, while the validation set was used for model generalization ability evaluation and predictive performance. Model performance was comprehensively evaluated using indicators such as the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy. To validate the clinical applicability of the models, we further plotted calibration curves, decision curves, PR curves, confusion matrices, and heatmaps, and utilized Shapley Additive Explanations (SHAP) to visualize variable importance, quantifying the contribution weight of each feature to the predictive outcome. The model with the best overall performance was ultimately selected as the predictive model for rectal cancer lymph node metastasis. Based on the independent risk factors screened through multivariate logistic regression analysis, a risk scoring system for rectal cancer lymph node metastasis was constructed. Using the regression coefficients (β values) of each variable as weights, relative weight calculations were performed with the variable having the smallest β value as the baseline, and the weights were standardized into integer scores to enhance clinical usability. Patients were stratified into low, medium, and high-risk tiers based on the total score for rapid assessment of lymph node metastasis risk. On this basis, this study developed a rectal cancer lymph node metastasis visualization online predictive calculator using the R language Shiny framework, based on the optimal model, which can output the

predictive probability of lymph node metastasis for patients in real time, intuitively displaying the impact of each factor on individual predictive results, and achieving clinical transformation of the model.

3 Results

3.1 Baseline Characteristics

A total of 500 enrolled patients were categorized into LNM-negative (n=176) and LNM-positive (n=324) cohorts according to postoperative pathological confirmation of nodal metastasis status. All participants were further randomly split into a training cohort (n=350) and an internal validation cohort (n=150) at a 7:3 ratio for subsequent model development and verification. The baseline clinicopathological characteristics and laboratory indicators of the two groups are shown in (Table 1). The analysis results showed multiple significant differences in demographic characteristics, tumor-related indicators, laboratory test indicators, and peripheral blood cell comprehensive score (PBCS): the age of the metastasis group was significantly higher than that of the non-metastasis group ($P < 0.001$); the tumor diameter was larger in the metastasis group, with a statistically significant difference compared to the non-metastasis group ($P < 0.001$); the distribution of histopathological classification ($P=0.002$), tumor cT stage ($P < 0.001$), and vascular invasion ($P < 0.001$) showed statistically significant differences between the two groups; in laboratory indicators, the levels of carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA19-9) in the metastasis group were significantly higher than those in the non-metastasis group ($P < 0.001$, $P=0.003$); furthermore, the PBCS score of the metastasis group was significantly higher than that of the non-metastasis group ($P < 0.001$). To ensure the reliability of subsequent model construction and validation, a baseline feature balance test was performed on the training set and validation set. The results showed that none of the above variables had statistically significant differences in distribution and levels between the training set and validation set (all $P > 0.05$), indicating that the random grouping method used in this study was scientifically effective, and the baseline data of the training set and validation set are balanced and comparable, making them reliable subjects for subsequent model construction and external validation.

3.2 Logistic Regression for Feature Selection in Model Development

Based on all clinical features, after stepwise retrospective screening, five risk factors were identified using univariate and multivariate logistic regression. Five independent predictors were identified as follows: age (OR = 1.03, 95% CI = 1.00-1.06, P = 0.024), tumor diameter (OR = 1.15, 95% CI = 1.01-1.31, P = 0.031), peripheral blood cell score (OR = 1.20, 95% CI = 1.01-1.42, P=0.040), carcinoembryonic antigen (OR = 1.12, 95% CI = 1.01-1.25, P=0.033), and tumor cT staging (OR = 1.78, 95% CI = 1.08-2.93, P=0.023) were identified as independent predictors of lymph node metastasis in rectal cancer patients. (Table 2) Based on these findings, machine learning models were constructed using the five independent predictors.

3.3 Performance of the Developed Models

In the training set, the Extreme Gradient Boosting (XGB) model demonstrated excellent predictive performance with an AUC of 0.873, 95% CI (0.836-0.909), and good prediction accuracy. The AUC values of the remaining seven models were as follows: Logistic Regression (LR) 0.681, 95% CI (0.618-0.743); Decision Tree (DT) 0.878, 95% CI (0.841-0.916); Random Forest (RF) 0.909, 95% CI (0.878-0.940); Support Vector Machine (SVM) 0.857, 95% CI (0.815-0.899); Multilayer Perceptron (MLP) 0.885, 95% CI (0.852-0.919); K-Nearest Neighbors (KNN) 0.877, 95% CI (0.842-0.912); LightGBM 0.891, 95% CI (0.856-0.925) (Figure 2A). In the validation set, the predictive performance of the models showed significant differences, with the XGB model still performing the best, achieving an AUC of 0.874, 95% CI (0.810-0.938), and outstanding generalization ability. The AUC values and 95% CIs of the remaining models were: LR 0.839 (0.764-0.914), DT 0.740 (0.660-0.821), RF

0.861 (0.793-0.930), SVM 0.804 (0.719-0.889), MLP 0.798 (0.715-0.882), KNN 0.814 (0.740-0.889), LGBM 0.739 (0.658-0.820) (Figure 2B). In the Decision Curve Analysis (DCA), Precision-Recall (PR) curves, and calibration curves of both the training and validation sets, the XGB model showed better consistency between predicted and actual metastatic probabilities compared to the other seven models, demonstrating higher clinical net benefit across all threshold ranges, indicating good clinical decision-making value and practical predictive ability (Figures 2C-H). This study used 10-fold cross-validation to robustly evaluate the performance of all machine learning algorithms, with results shown in Figure 3. The average AUC values of the models were as follows: LR 0.666, DT 0.806, RF 0.812, XGB 0.808, SVM 0.816, MLP 0.833, KNN 0.813, LGBM 0.828. Figure 4 shows the multi-series estimation line graph of the training dataset, intuitively presenting the trend of indicator changes in the training set for each model. Figure 5 visualizes the confusion matrix, revealing that the XGB model performed stably in both positive and negative sample detection, with significantly better sensitivity than the other models. Figure 6 provides a visual comparison of the comprehensive predictive performance of each model, with results indicating that the XGB model was at a high level in key metrics such as accuracy, precision, and sensitivity. Comprehensive evaluation across ROC, DCA, PR, and calibration curves showed that MLP and LightGBM achieved comparable cross-validation AUC to XGB, whereas RF exhibited superior AUC in the training set, each model had its advantages. However, the XGB model exhibited the most stable performance between the training and validation sets with the lowest risk

of overfitting. DCA curves showed higher clinical net benefit, calibration curves indicated the best consistency between predicted probabilities and true risks, and PR curves demonstrated superior area under the curve, making it more suitable for this study. Considering stability, clinical applicability, and predictive accuracy, the XGB model showed the most balanced overall performance, and thus was selected for constructing the predictive model.

3.4 Variable Importance

The importance of variables in each ML algorithm was quantified (Figure 7). In the XGB model, the importance of variables was ranked as follows: tumor diameter, CEA, age, tumor cT stage, PBCS.

3.5 Model Interpretability

SHAP analysis was performed to quantify the contribution of each predictive variable. Figure 8 visually presents the importance ranking of each feature, where each point represents a feature, and the distribution of points reflects the extent of the feature's impact on the model. The vertical position of each point corresponds to the SHAP value of the feature, which indicates the magnitude of the feature's contribution to the model's output. A positive value indicates that the feature promotes a positive predictive outcome, while a negative value indicates an inhibitory effect. In terms of color, orange represents high feature values, and purple represents low feature values. The darker the color, the greater the feature's impact on the predictive outcome. From the figure, it can be observed that higher values of tumor diameter, CEA, age, tumor cT stage, and PBCS are associated with an increased risk of rectal cancer lymph node metastasis.

3.6 Construction of the Clinical Risk Scoring System

To further enhance the clinical interpretability and practicality of the model, this study constructed a risk scoring system for lymph node metastasis in rectal cancer based on 5 independent risk factors selected by multivariate Logistic regression (Table 3). Using the regression coefficients (β values) of each variable as weights, the age with the smallest β value was set as the reference variable and standardized into integer scores based on relative weights, with a total score range of 0–26 points. Patients were categorized into three risk groups based on the total score: low-risk (0–5 points), medium-risk (6–12 points), and high-risk (≥13 points), with corresponding theoretical lymph node metastasis probabilities of <20%, 20%–50%, and >50%, respectively. Stratified analysis showed that the lymph node metastasis rate of patients increased stepwise with the increase in risk level, and the scoring system showed high consistency with the predictive performance of the XGB model, making it a simplified clinical alternative tool for complex machine learning models to achieve rapid preoperative quantitative evaluation.

3.7 Construction and Application of an Online Calculator

To further enhance the clinical accessibility and utility of the model, this study developed an online calculator for predicting lymph node metastasis in rectal cancer (https://xjkjzykd.shinyapps.io/Transfer/) based on the optimal XGB model and the aforementioned risk scoring system. The tool uses five independent risk factors—age, tumor diameter, CEA, PBCS, and tumor cT stage—as input indicators, enabling real-time individualized risk probability calculation through a simple interactive

interface. Clinicians can quickly obtain lymph node metastasis risk predictions by inputting preoperative patient data, which corroborates and complements the risk scoring system constructed in this study. The online calculator does not require specialized statistical software, is easy to operate, and provides intuitive results, effectively addressing the limitations of machine learning models and achieving a complete transformation from a predictive model to a clinical utility tool.

4 Discussion

Colorectal cancer (CRC) is the third most common malignant tumor globally and the second leading cause of cancer-related deaths (18). Despite continuous advancements in surgery, chemotherapy, radiotherapy, and immunotherapy, rapid tumor progression and widespread metastasis still result in a 5-year survival rate of less than 20% for patients (19). Among these, lymph node metastasis directly influences the treatment strategy and prognosis of colorectal cancer patients. Although computer tomography (CT) and magnetic resonance imaging (MRI) are commonly used for diagnosing lymph node metastasis, their ability to identify micro-metastatic lesions is limited. Therefore, exploring more convenient and cost-effective predictive factors for lymph node metastasis is particularly important in the diagnostic and treatment process. This study involves screening based on a combination of general patient information, imaging data, and serum tumor markers, with the aim of incorporating machine learning models for training.

This study places high emphasis on the scientific rigor and strictness of the research methodology, adhering strictly to the predefined inclusion and exclusion criteria for

subject selection to ensure high homogeneity among enrolled patients in terms of rectal cancer lymph node metastasis outcomes, effectively avoiding research bias caused by case heterogeneity. Compared to previous similar studies, this study achieves more precise retrospective analysis of clinical data from both groups through rigorous enrollment screening, aiming to identify key driving factors affecting lymph node metastasis (LNM) and provide clinicians at the forefront with more optimized and precise treatment strategies to achieve the best treatment outcomes.

We utilized a new type of artificial intelligence—machine learning—to predict lymph node metastasis of rectal cancer in patients after RC surgery. In terms of model construction and validation, this study employed eight classic machine learning algorithms to predict the likelihood of rectal cancer lymph node metastasis. The model parameters were optimized through 10-fold cross-validation, and comprehensive evaluation of model performance was conducted using multi-dimensional indicators such as AUC, calibration curves, DCA curves, PR curves, and confusion matrices. Ultimately, it was found that the XGB model performed the best, with an AUC of 0.873 in the training set and 0.874 in the validation set. The calibration curve also showed good consistency between predicted probabilities and actual observed probabilities, while the DCA curve confirmed its high clinical net benefit across all threshold ranges. Compared to traditional nomogram models, the XGB model adopted in this study exhibits stronger nonlinear fitting capabilities, enabling better exploration of the complex relationship between clinical indicators and lymph node metastasis. Additionally, SHAP analysis was used

to achieve model interpretability, intuitively demonstrating the contribution degree of each variable to the prediction results. Through this research, we identified age, tumor diameter, CEA, PBCS, and tumor cT stage as independent predictive factors for lymph node metastasis in rectal cancer patients.

Multiple studies both domestically and internationally have confirmed that the levels of inflammatory response markers are inextricably linked to the prognosis of CRC patients. Changes in these marker levels reflect the complex interactions between tumor-related systemic inflammation, the tumor microenvironment, and host immunity. Incorporating these markers into clinical practice can improve the accuracy of prognosis prediction and provide references for treatment decisions (20). The academic teams of Lin N and Acikgoz O, respectively, pointed out in their studies that elevated NLR and PLR are significantly associated with poor prognosis in colorectal cancer (21, 22), which aligns with the results of this study. Interestingly, the PAR analyzed in this study was identified by Wang K's team in 2025 as a reliable prognostic predictor for colorectal cancer (23). Meanwhile, the only negatively correlated inflammatory marker in this study, HPR, was also highlighted by Zeynelgil E's team in the same year: HPR has an important relationship with the prognosis of CRC patients (24). The core innovation of this study lies in constructing a PBCS scoring system based on four peripheral blood cell ratios—PLR, NLR, PAR, and HPR—and incorporating it as an independent variable into the machine learning model. Existing research has primarily focused on the clinical value of single blood cell ratios, lacking exploration of multi-indicator combined applications (25). This

study determined the optimal cutoff values for each indicator using ROC curves and developed a 0-4 point PBCS scoring system to achieve quantitative integration of multiple inflammatory markers. Multivariate regression results confirmed that the PBCS score is an independent risk factor for lymph node metastasis, further validating the clinical value of this scoring system and addressing the limitations of single-indicator prediction.

Furthermore, we did not overlook serum tumor markers that share the same origin as inflammatory markers. The SHAP graph clearly shows that CEA is the second most critical variable in our study. Scholars both domestically and internationally have previously highlighted that preoperative CEA is an important indicator of distant metastasis in rectal cancer (26, 27). Although CEA is a broad-spectrum tumor marker and cannot serve as a specific diagnostic indicator for a particular malignant tumor, it still holds significant clinical value in differential diagnosis of malignant tumors, disease monitoring, and evaluation of treatment efficacy (28). Therefore, we believe that rectal cancer patients with higher preoperative CEA levels require more optimized postoperative treatment plans and should be vigilant about the possibility of lymph node metastasis.

In univariate and multivariate analyses, we found that tumor diameter is also one of the independent risk factors, and its importance ranks first in the model of this study. In previous multiple studies, the size of the tumor as an independent risk factor for cancer metastasis has been repeatedly studied and validated by scholars (29-31). Luo et al. also found that an increase in rectal tumor volume is associated with an

increased risk of distant metastasis (32). The results of multivariate logistic regression and machine learning models in this study are consistent with these findings. The larger the tumor, the longer the tumor growth time, and the higher the evolutionary degree of tumor cells, which may further invade surrounding soft tissues, potentially explaining the relationship between tumor size and distant metastasis.

Interestingly, our research revealed that age and tumor cT stage are also among the influencing factors. Although some scholars have pointed out that the positive rate of lymph node metastasis (LNM) in young patients with colorectal cancer (CRC) is higher than in elderly patients, suggesting that the tumor has rapid growth, strong invasion, and a tendency to metastasize (33), we believe that CRC is primarily an elderly disease, mainly affecting the elderly population. This is because as one ages, the immune function of the body gradually declines, while the immune escape ability of tumor cells is further enhanced. Additionally, changes in the microenvironment of elderly individuals are more conducive to the invasion and metastasis of tumor cells. These factors collectively contribute to an increased risk of lymph node metastasis. Building on this foundation, our study further validates these findings, providing new empirical evidence for the disease progression mechanisms in elderly tumor patients. This has significant guiding implications for the development and optimization of clinical diagnostic and treatment strategies. Regarding tumor cT stage, we referred to the Colorectal Cancer Treatment Guidelines published by the Japanese Society for Cancer of the Colon and Rectum (JSCCR) in 2024 (34). Scholars in the document noted that cases with tumor invasion depth exceeding 1000 μm have an increasing

risk of lymph node metastasis as the invasion depth deepens. Meanwhile, foreign scholar Han J suggested that patients with an invasion depth exceeding 1900 μm have a 19.3% risk of LNM (35). According to our research results, tumor cT stage is also an important indicator for assessing the occurrence of LNM and prognosis in CRC. Through preoperative endoscopic ultrasound, high-resolution MRI, enhanced CT, colonoscopy, and biopsies, we preliminarily determine the tumor invasion depth, classify them into two groups—T1+T2 (superficial invasion group) and T3+T4 (deep invasion group)—according to the reference standards of the American Joint Committee on Cancer (AJCC), and include them in the model for training.

To further enhance the clinical applicability and accessibility of the model, this study constructed a simple clinical risk scoring system based on the aforementioned independent risk factors. This scoring system includes only five preoperative easily obtainable indicators, enabling rapid stratification into low, medium, and high-risk categories for quick assessment of lymph node metastasis risk. Additionally, this study developed an online predictive calculator based on the Shiny framework, which can output personalized prediction probabilities and SHAP explanations in real time, achieving clinical translation of the model and significantly improving the clinical translational value of this research.

Although the predictive model developed in our study demonstrates good accuracy, aiding frontline clinicians in rapidly and conveniently assessing the likelihood of lymph node metastasis in rectal cancer patients, this study still has certain limitations that require continuous optimization and improvement in subsequent clinical practice.

First, the results of some baseline variables differ from previous studies and should be interpreted with caution: in this study, neural invasion and lymph node metastasis did not reach statistical significance (P=0.0614), which may be related to sample size, selection bias in single-center retrospective data, and the timing of assessment, and requires further validation by expanding the sample size. Additionally, smoking history showed a negative correlation with lymph node metastasis (32.1% in the metastasis group vs. 40.9% in the non-metastasis group), which is inconsistent with some previous study results. This may be related to confounding factors such as the baseline characteristics of the study population and health behavior bias, and this result still needs to be further confirmed by prospective, multicenter studies. Second, this study is a single-center retrospective cohort study, and data missingness and selection bias cannot be completely avoided, limiting the applicability of the model. Although the sample size included in this study reached 500, further validation of the conclusions of this study still needs to be conducted through multicenter, large-sample prospective cohort studies. Additionally, PBCS is constructed using equal weights and does not perform weighted optimization on the indicators, which cannot accurately reflect the predictive contribution differences of different inflammatory markers to lymph node metastasis. Regarding inflammatory markers, it is possible that further research by scholars may discover more and more accurate predictive factors and incorporate them into model training. Additionally, the online calculator is currently based on the data of this study and its external applicability and clinical application value still need to be further evaluated.

5 Conclusion

In summary, we developed a machine learning-based predictive model for LNM in RC patients. During model training, age, tumor diameter, tumor cT staging, CEA, and PBCS were identified as independent predictors of rectal cancer lymph node metastasis. The model trained based on these factors has good predictive performance and can provide clinicians with important references for formulating precise and individualized treatment plans. At the same time, the clinical risk scoring system constructed in this study is easy to operate and has clear stratification; The developed online prediction calculator enables real-time personalized evaluation, effectively promoting the clinical translation of the model and providing practical tools for precise preoperative stratification and personalized diagnosis and treatment of rectal cancer patients.

Statements

Data availability statement

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

Ethics statement

This retrospective study was approved by the Ethics Committee of Jinzhou Central Hospital of Jinzhou Medical University (Ethics Approval No. 2025-LW-034). The study was conducted in accordance with the 1964 Declaration of Helsinki. Informed consent was waived by the ethics committee due to the retrospective nature of this study.

Author contributions

JX: Writing – original draft, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing.

BC: Writing – review & editing, Data curation, Formal analysis, Investigation.

WZ: Writing – review & editing Conceptualization, Data curation, Investigation, Methodology.

CG: Writing – review & editing, Data curation, Investigation.

ZY: Writing – review & editing, Data curation, Investigation.

WZ (Weibin Zhang): Writing – review & editing, Conceptualization, Methodology, Supervision.

Funding

This study was supported by the Natural Science Foundation of Liaoning Province (Grant No. 2025-MS-371).

Acknowledgments

Not applicable.

Consent for publication

Not applicable.

Conflict of interest

The authors declare that they have no competing interests.

Generative AI statement

The authors declared that generative AI was not used in the creation of this manuscript.

Abbreviations

Abbreviations: RC, Rectal cancer; LNM, Lymph node metastasis; CEA, Carcinoembryonic antigen; PBCS, Peripheral blood cell score; PLR, Platelet-to-lymphocyte ratio; NLR, Neutrophil-to-lymphocyte ratio; ROC, Receiver Operating Characteristic; SHAP, SHapley Additive Explanations; AUC, Area Under Curve; PR, Precision-Recall Curve; PAR, Platelet-to-albumin ratio; ML, Machine Learning; HPR, Hemoglobin-platelet ratio; WHO, World Health Organization; AFP, Alpha fetoprotein; ADA, American Diabetes Association; OGTT, Oral Glucose Tolerance Test; AJCC, American Joint Committee on Cancer; JSCCR, Japanese Society for Cancer of the Colon and Rectum; NCCN, National Comprehensive Cancer Network; ASCO, American Society of Clinical Oncology.

References

1 GBD 2019 Colorectal Cancer Collaborators. Global, regional, and national burden of colorectal cancer and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Gastroenterol Hepatol* (2022) 7(7):627-47. doi: 10.1016/s2468-1253(22)00044-9

2 Akimoto N, Ugai T, Zhong R, Hamada T, Fujiyoshi K, Giannakis M, et al. Rising incidence of early-onset colorectal cancer — a call to action. *Nat Rev Clin Oncol* (2021) 18(4):230-43. doi: 10.1038/s41571-020-00445-1

3 Ganesh K, Stadler ZK, Cercek A, Mendelsohn RB, Shia J, Segal NH, et al. Immunotherapy in colorectal cancer: rationale, challenges and potential. *Nat Rev Gastroenterol Hepatol* (2019) 16(6):361-75. doi: 10.1038/s41575-019-0126-x

4 Cai MZ, He HD, Hong SJ, Weng JM. Synergistic diagnostic value of circulating tumor cells and tumor markers CEA/CA19-9 in colorectal cancer. *Scand J Gastroentero* (2023) 58(1):54-60. doi: 10.1080/00365521.2022.2106152

5 Ren F, Weng WW, Zhang QY, Tan C, Xu MD, Zhang M, et al. Clinicopathological features and prognosis of AFP-producing colorectal cancer: a single-center analysis of 20 cases. *Cancer Manag Res* (2019) 11:4557-67. doi: 10.2147/CMAR.S196919

6 Björkman K, Mustonen H, Kaprio T, Kekki H, Pettersson K, Haglund C, et al. CA125: A superior prognostic biomarker for colorectal cancer compared to CEA, CA19-9 or CA242. *Tumor Biol* (2021) 43:57-70. doi: 10.3233/TUB-200069

7 Fernandes JV, Cobucci RNO, Jatobá CAN, de Medeiros Fernandes TAA, de Azevedo JWV, de Araújo JMG. The Role of the Mediators of Inflammation in Cancer Development. *Pathol Oncol Res* (2015) 21(3):527-34. doi: 10.1007/s12253-015-9913-z

8 Schlesinger M. Role of platelets and platelet receptors in cancer metastasis. *J Hematol Oncol* (2018) 11(1):125. doi: 10.1186/s13045-018-0669-2

9 St. Paul M, Ohashi PS. The roles of $CD8^+$ T cell subsets in antitumor immunity. *Trends Cell Biol* (2020) 30(9):695-704. doi: 10.1016/j.tcb.2020.06.003

10 Mizuno R, Kawada K, Itatani Y, Ogawa R, Kiyasu Y, Sakai Y. The role of tumor-associated neutrophils in colorectal cancer. *Int J Mol Sci* (2019) 20(3):529. doi: 10.3390/ijms20030529

11 Yamamoto T, Kawada K, Obama K. Inflammation-related biomarkers for the prediction of prognosis in colorectal cancer patients. *Int J Mol Sci* (2021) 22(15):8002. doi: 10.3390/ijms22158002

12 Cupp MA, Cariolou M, Tzoulaki I, Aune D, Evangelou E, Berlanga-Taylor AJ. Neutrophil to lymphocyte ratio and cancer prognosis: an umbrella review of systematic reviews and meta-analyses of observational studies. *BMC Med* (2020) 18(1):360. doi: 10.1186/s12916-020-01817-1

13 Xiao ZL, Wang XX, Chen XX, Zhou JW, Zhu HT, Zhang JN, et al. Prognostic role of preoperative inflammatory markers in postoperative patients with colorectal cancer. *Front Oncol* (2023) 13:1064343. doi: 10.3389/fonc.2023.1064343

14 Mo CJ, Hu ZJ, Qin SZ, Chen HP, Huang L, Li S, et al. Diagnostic value of platelet-lymphocyte ratio and hemoglobin-platelet ratio in patients with rectal cancer. *J Clin Lab Anal* (2020) 34(4):e23153. doi: 10.1002/jcla.23153

15 Făgărășan V, Bințințan V, Seicean R, Caziuc A, Ailioaie R, Făgărășan G, et al. Lymphocyte-to-monocyte, platelet-to-albumin and platelet-to-lymphocyte ratios as prognostic biomarkers for neoadjuvant treatment response in rectal cancer patients. *Surg Oncol* (2024) 56:102126. doi: 10.1016/j.suronc.2024.102126

16 Guo PY, Hu XH, Li BK, Lu T, Liu JM, Wang CY, et al. Peripheral blood cell count composite score as a prognostic factor in patients with colorectal cancer. *Chin J Gastrointest Surg* (2024) 27:953-65. doi: 10.3760/cma.j.cn441530-20231029-00151

17 Sakamoto T, Goto T, Fujiogi M, Kawarai Lefor A. Machine learning in gastrointestinal surgery. *Surg Today* (2022) 52(7):995-1007. doi: 10.1007/s00595-021-02380-9

18 Sung H, Ferlay J, Siegel R, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71(3):209-49. doi: 10.3322/caac.21660

19 Zheng RS, Zhang SW, Zeng HM, Wang SM, Sun KX, Chen R, et al. Cancer incidence and mortality in China, 2016. *J Natl Cancer Cent* (2022) 2(1):1-9. doi: 10.1016/j.jncc.2022.02.002

20 Kim M, Son IT, Oh BY. Inflammatory response markers as predictors of colorectal cancer prognosis. *Ewha Med J* (2023) 46(Suppl 1):e24. doi: 10.12771/emj.2023.e24

21 Lin NP, Li JR, Yao XQ, Zhang X, Liu GZ, Zhang ZB, et al. Prognostic value of neutrophil-to-lymphocyte ratio in colorectal cancer liver metastasis: A meta-analysis of results from multivariate analysis. *Int J Surg* (2022) 107:106959. doi: 10.1016/j.ijsu.2022.106959

22 Acikgoz O, Cakan B, Demir T, Bilici A, Oven BB, Hamdard J, et al. Platelet to lymphocyte ratio is associated with tumor localization and outcomes in

metastatic colorectal cancer. *Medicine* (2021) 100(44):e27712. doi: 10.1097/md.0000000000027712

23 Wang K, Li KJ, Zhang ZY, Zeng XY, Wu ZM, Zhang BX, et al. Combined preoperative platelet-albumin ratio and cancer inflammation prognostic index predicts prognosis in colorectal cancer: a retrospective study. *Sci Rep* (2025) 15(1):29500. doi: 10.1038/s41598-025-15309-w

24 Zeynelgil E, Duzkopru Y, Kocanoglu A, Karakaya S. Prognostic value of hemoglobin-to-red cell distribution width ratio and inflammation markers in colorectal cancer. *World J Gastrointest Oncol* (2025) 17(6):106603. doi: 10.4251/wjgo.v17.i6.106603

25 Shen X, Xiang MY, Tang JD, Xiong GR, Zhang K, Xia TR, et al. Evaluation of peripheral blood inflammation indexes as prognostic markers for colorectal cancer metastasis. *Sci Rep* (2024) 14(1):20489. doi: 10.1038/s41598-024-68150-y

26 Tarantino I, Warschkow R, Schmied BM, Güller U, Mieth M, Cerny T, et al. Predictive value of CEA for survival in stage i rectal cancer: a population-based propensity score-matched analysis. *J Gastrointest Surg* (2016) 20(6):1213-22. doi: 10.1007/s11605-016-3137-8

27 Zhang BB, Sun ZZ, Song M, Ma SQ, Tian YJ, Kong QF. Ultrasound/CT combined with serum CEA/CA19.9 in the diagnosis and prognosis of rectal cancer. *J BUON* (2018) 23:592-7.

28 Lakemeyer L, Sander S, Wittau M, Henne-Bruns D, Kornmann M, Lemke J. Diagnostic and prognostic value of CEA and CA19-9 in colorectal cancer. *Diseases* (2021) 9(1):21. doi: 10.3390/diseases9010021

29 Xu C, Zhou Q, Liu WC, Li WL, Dong ST, Li WY, et al. Dynamic predictive models with visualized machine learning for assessing the risk of lung metastasis in kidney cancer patients. *J Oncol* (2022) 2022:5798602. doi: 10.1155/2022/5798602.

30 Xiao RT, Liu C, He W, Ma LL. Prognostic factors and a nomogram predicting overall survival and cancer-specific survival for patients with collecting duct

renal cell carcinoma. *Biomed Res Int* (2021) 2021(1):6736008. doi: 10.1155/2021/6736008.

31 Tang C, Xu JM, Lin MB, Qiu SX, Wang H, Zuo XM, et al. Risk factors for distant metastasis in T3 T4 rectal cancer. *Clin Med Insights Oncol* (2024) 18:1-12. doi: 10.1177/11795549241227423.

32 Luo DK, Shan ZZ, Liu Q, Cai SJ, Ma YL, Li QG, et al. The correlation between tumor size, lymph node status, distant metastases and mortality in rectal cancer patients without neoadjuvant therapy. *J Cancer* (2021) 12(6):1616-22. doi: 10.7150/jca.52165.

33 Zhang QW, Sun LC, Tang CT, Liang Q, Zhou YY, Chen HM, et al. Inverse association of age with risk of lymph node metastasis in superficial colorectal cancer: A large population-based study. *Oncologist* (2020) 25(6):e920-e7. doi: 10.1634/theoncologist.2019-0815.

34 Kinugasa Y, Uehara K, Yamaguchi K, Saito Y, Murofushi K, Sugai T, et al. Japanese Society for Cancer of the Colon and Rectum (JSCCR) guidelines 2024 for the treatment of colorectal cancer. *Int J Clin Oncol* (2025) 30(12):2410-63. doi: 10.1007/s10147-025-02899-8.

35 Han J, Hur H, Min BS, Lee KY, Kim NK. Predictive factors for lymph node metastasis in submucosal invasive colorectal carcinoma: A new proposal of depth of invasion for radical surgery. *World J Surg* (2018) 42(8):2635-41. doi: 10.1007/s00268-018-4482-4.

36 Hughes, G., Youden's index and the weight of evidence. Methods Inf Med, 2015. 54(2): p. 198-9.

Figure legends

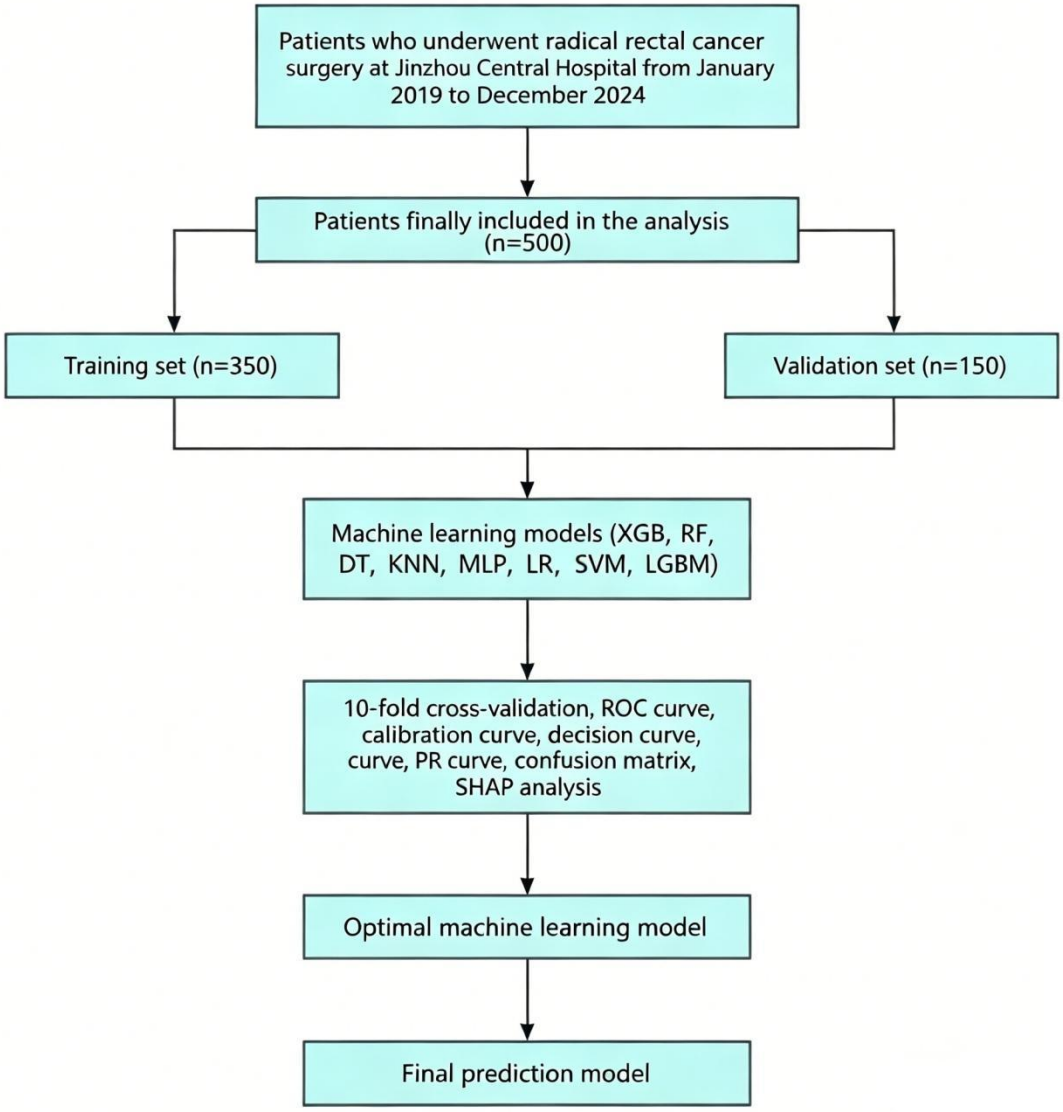


Figure 1. Technology Roadmap.

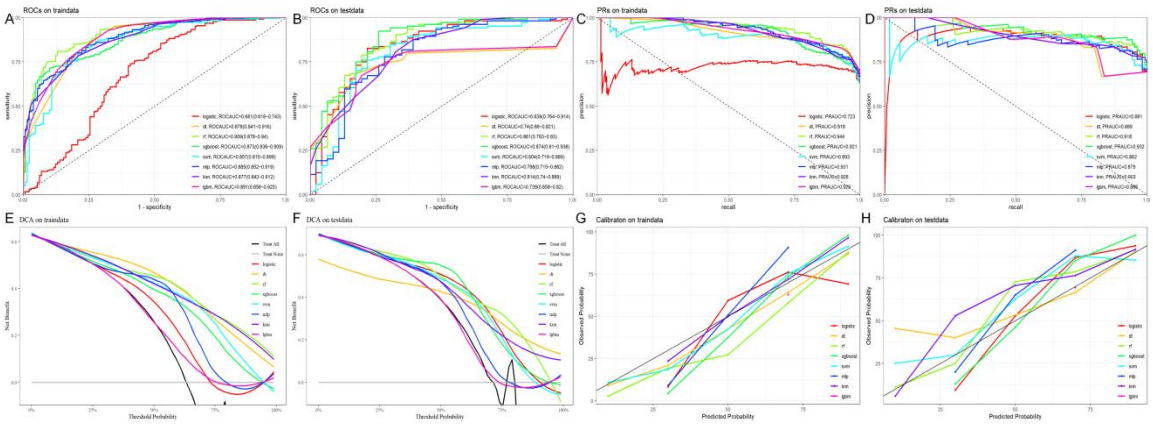


Figure 2. (A) ROC curve of the training set (B) ROC curve of the validation set (C) PR curve of the training set (D) PR curve of the validation set (E)

DCA curve of the training set (F) DCA curve of the validation set (G) Calibration curve of the training set (H) Calibration curve of the validation set.

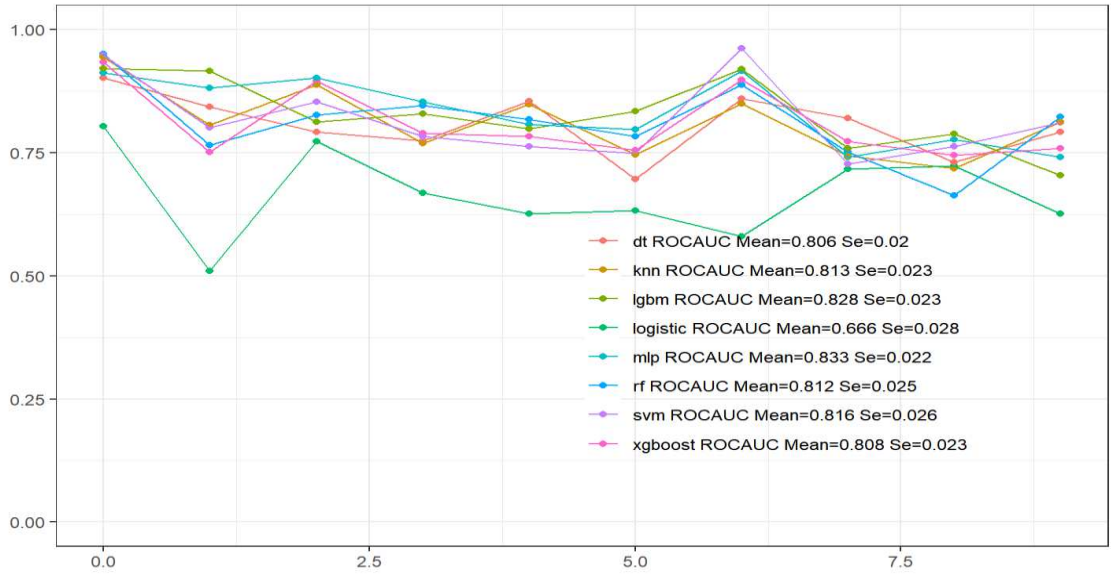


Figure 3. Results of 10-fold cross-validation for different machine models in the training set.

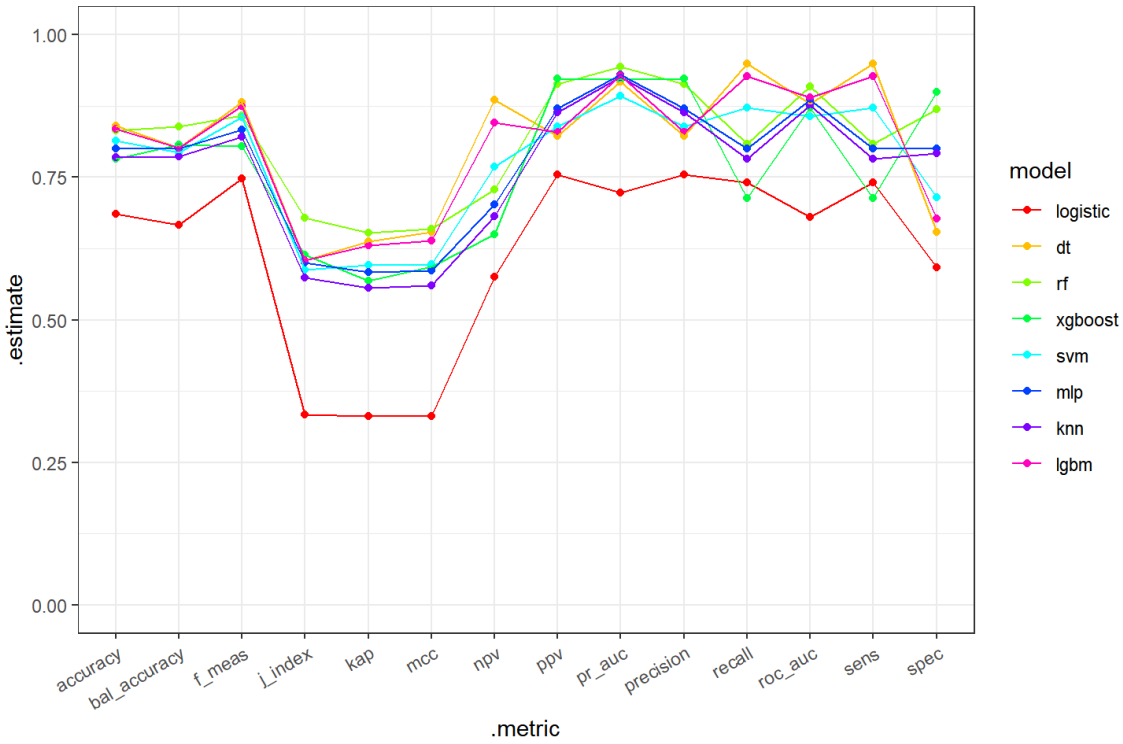


Figure 4. Line chart of multi-series estimated values in the training set.

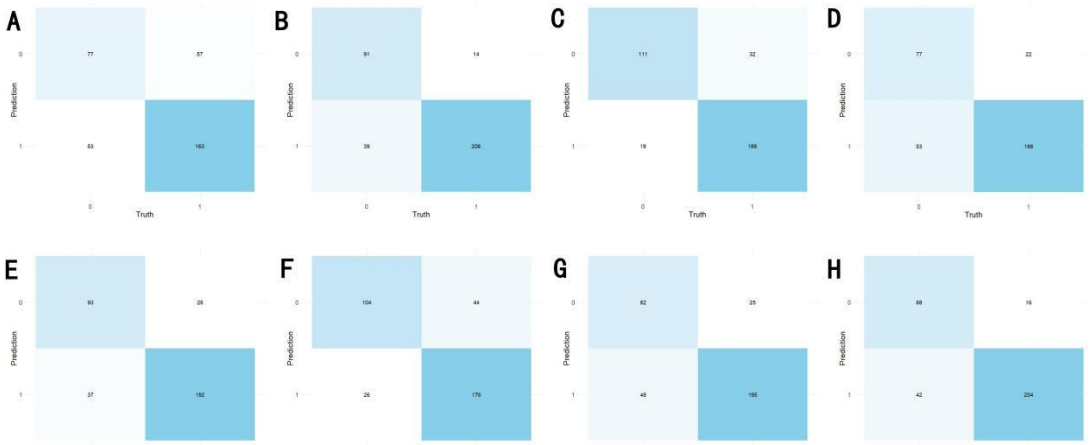


Figure 5. Confusion matrices of various machine learning models (A: LR, B: DT, C: RF, D: XGB, E: SVM, F: MLP, G: KNN, H: LGBM), providing a visual comparison of the detection performance of each model for positive and negative samples.

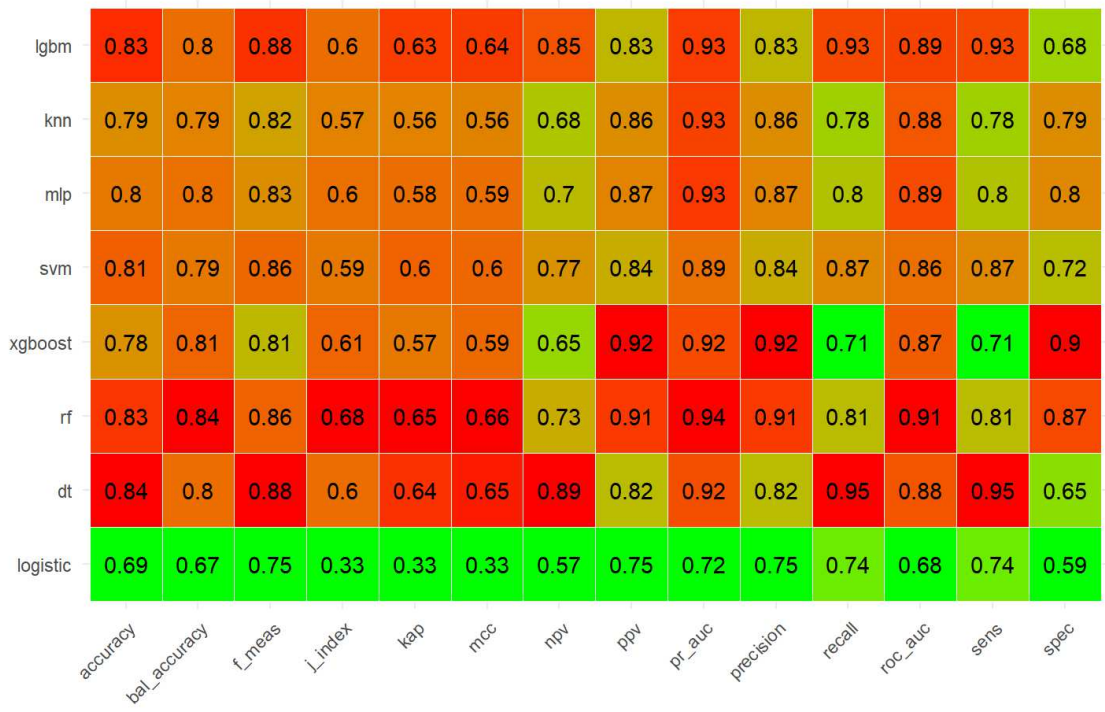


Figure 6. Predictive performance of different models.

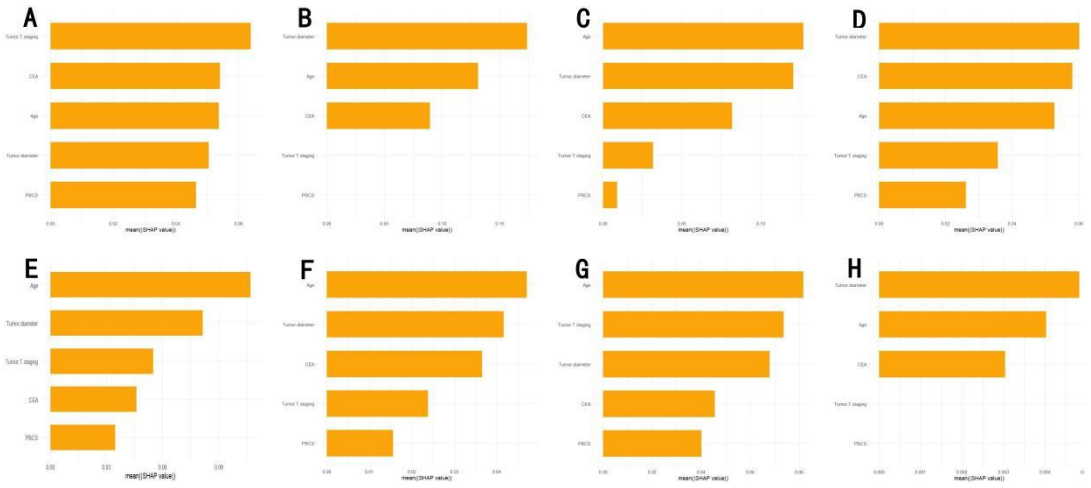


Figure 7. Importance ranking of predictive variables for each machine learning model (A: LR, B: DT, C: RF, D: XGB, E: SVM, F: MLP, G: KNN, H: LGBM), where bar length represents the contribution of the variable to the model's prediction.

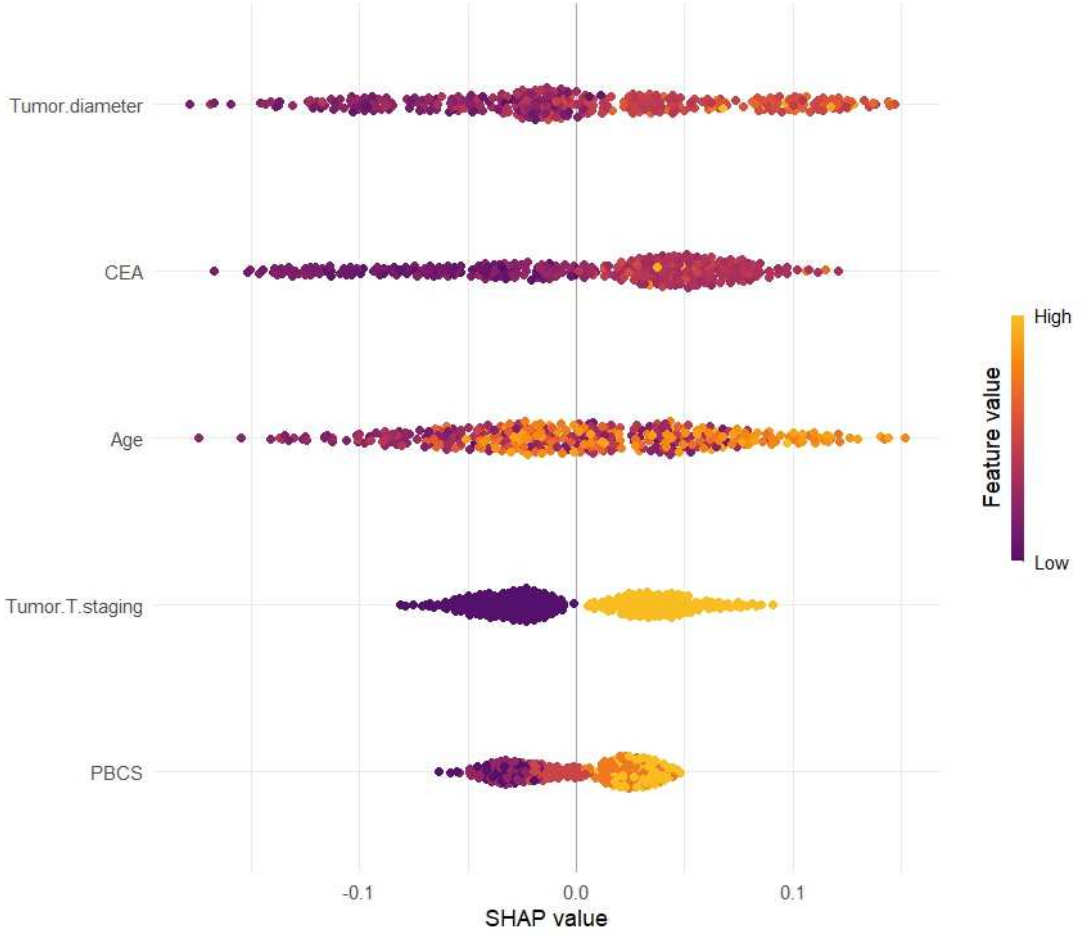
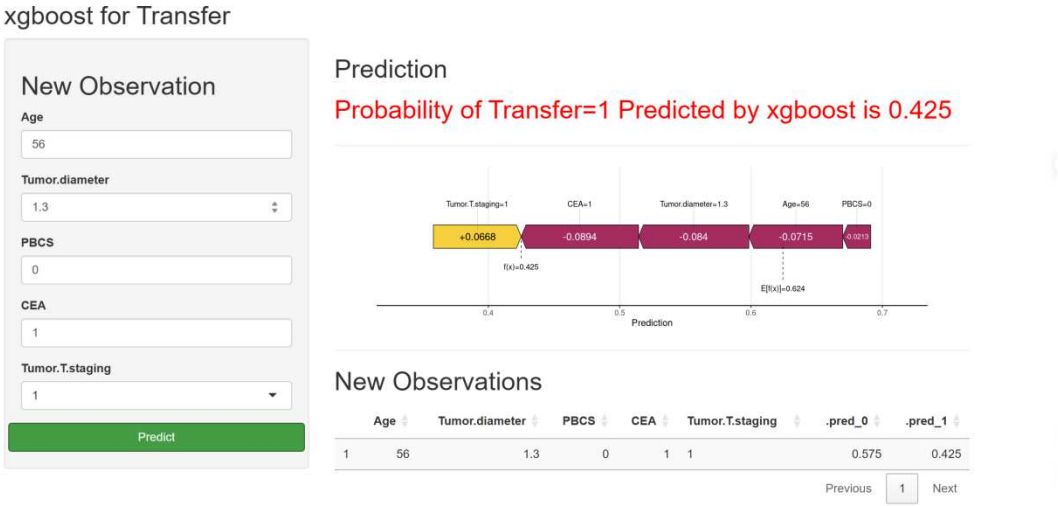


Figure 8. SHAP plot.



730

731 Figure 9. Screenshot of the online calculator

732

733 **Tables**

734 Table 1 Comparison of baseline characteristics in the original dataset

Variable	Normal group	Metastatic group	*P*	Training set	Validation set	*P*
	N = 176	N = 324		N = 350	N = 150	
Age	69(15.25)	76(15)	<0.001**	74(16.75)	72(14)	0.541
Sex			0.504			0.466
Male	72(40.9%)	144(44.4%)		147(42.0%)	69(46.0%)	
Female	104(59.1%)	180(55.6%)		203(58.0%)	81(54.0%)	
Smoking	72(40.9%)	104(32.1%)	0.061	127(36.3%)	49(32.7%)	0.500
Hypertension	59(33.5%)	123(38.0%)	0.374	123(35.1%)	59(39.3%)	0.429
Drinking	57(32.39%)	99(30.56%)	0.748	113(32.3%)	43(28.7%)	0.487
Family history	26(14.77%)	54(16.67%)	0.672	59(16.86%)	21(14.00%)	0.506
Diabetes	32(18.18%)	52(16.05%)	0.629	54(15.4%)	30(20.0%)	0.262
tumor diameter	4(2.8)	5.95(3.3)	<0.001**	5.5(2.8)	5.65(4.2)	0.476
Histopathological classification			0.002*			0.509
Poorly differentiated	21(11.93%)	73(22.53%)		63(18.00%)	31(20.67%)	
Moderately differentiated	87(49.43%)	166(51.23%)		183(52.29%)	70(46.67%)	
Well differentiated	68(38.64%)	85(26.23%)		104(29.71%)	49(32.67%)	
Gross tumor classification			0.268			0.173
Ulcerative type	146(83.0%)	282(87.0%)		305(87.1%)	123(82.0%)	
Protrudin	30(17.0%)	42(13.0%)		45(12.9%)	27(18.0%)	

g type						
Tumor cT staging	54(30.7%)	180(55.6%)	<0.001**	163(46.57%)	71(47.33%)	0.953
Nerve invasion	28(15.9%)	76(23.5%)	0.061	72(20.57%)	32(21.33%)	0.943
Vascular invasion	22(12.5%)	86(26.5%)	<0.001**	72(20.6%)	36(24.0%)	0.462
Peripheral blood cell score	2(2)	3(3)	<0.001**	3(2)	3(1)	0.237
Hemoglobin	131(28.75)	128(32.5)	0.366	129(31.75)	131(30)	0.731
Fibrinogen	3.53(0.97)	3.66(1.12)	0.126	3.64(1.11)	3.59(0.95)	0.104
Carcinoembryonic antigen	4.27(3.00)	6.1(2.9)	<0.001**	5.8(3.04)	5.55(3.36)	0.962
Carbohydrate Antigen 19-9	10.55(15.43)	14.9(26.83)	0.003*	13.9(22.05)	11.9(18.6)	0.257
Sialylated antigen 72-4	2.4(4.22)	3(4.32)	0.147	2.79(4.45)	2.5(3.87)	0.244
Carbohydrate Antigen 125	6.5(14.38)	5.34(10.18)	0.234	5.34(11.2)	6.5(13.35)	0.503
Microsatellite instability	176(100%)	316(97.53%)	0.062	345(98.57%)	147(98.00%)	0.679
Mismatch repair protein MSH6	176(100%)	320(98.77%)	0.305	349(99.71%)	147(98.00%)	0.086
Mismatch repair protein MSH2	175(99.43%)	319(98.46%)	0.444	348(99.43%)	146(97.33%)	0.067
Mismatch repair protein MLH1	175(99.43%)	319(98.46%)	0.419	348(99.43%)	146(97.33%)	0.081

Mismatch repair protein PMS2	176(100%)	320(98.77%)	0.309	348(99.43%)	148(98.67%)	0.599

Note: * indicates $P < 0.05$, ** indicates $P < 0.001$.

Table 2 Multivariate logistic regression analysis

Variable	*OR* value	95%*CI*	*P*
Age	1.03	1.00-1.06	0.024
tumor diameter	1.15	1.01-1.31	0.031
Peripheral blood cell score	1.20	1.01-1.42	0.040
Carcinoembryonic antigen	1.12	1.01-1.25	0.033
Tumor cT staging	1.78	1.08-2.93	0.023

Table 3 Risk scoring system for lymph node metastasis in patients with Rectal Cancer

Variable	Grouping	Assignment (Score)
Age (years)	＜70	0
	≥70	1
tumor diameter (cm)	＜4	0
	4–5.9	3
	≥6	5
Peripheral blood cell score	0–1	0
	2–3	3
	4	6
Carcinoembryonic antigen (ng/ml)	＜5	0
	5–10	2
	＞10	4
Tumor cT staging	T1+T2	0
	T3+T4	10

Note: The total score ranges from 0 to 26; Low-risk: 0 - 5 points, Moderate-risk: 6 - 12 points,

High-risk: ≥ 13 points.

Supplementary Files

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

- SupplementaryFileS1PBCS.docx