Pharmacogenomics and Personalized Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Construction and Validation of Multi-Omics Predictive Models for Colorectal Cancer Using Machine-Learning Approaches

Zhenhuan Lu[1,*], Xiaowen Li[1,*], Zhiping Liang[1], Xiaocong Zhang[1], Yiyan Tan[1], Yinglan Kuang[2], Kang Li[1], Xiaofeng Zhu[1]

[1]Department of Gastrointestinal Surgery, Yuebei People's Hospital, Shaoguan City, Guangdong Province, People's Republic of China; [2]A.I. R&D Center, Zhuhai Hengqin Sanmed Aitech Ltd., Zhuhai City, Guangdong Province, People's Republic of China

*These authors contributed equally to this work

Correspondence: Xiaofeng Zhu; Kang Li, Department of Gastrointestinal Surgery, Yuebei People's Hospital, No. 133 Huimin South Road, Wujiang District, Shaoguan City, Guangdong Province, 512000, People's Republic of China, Tel +86-07518101207, Email Zhuxiaofengmn@163.com; Lykangerkl@21cn.com

Objective: To construct and externally validate a multi-omics nomogram that uses only routine clinicopathological variables to predict tumor mutational burden (TMB), microsatellite instability (MSI), NTRK/PIK3CA mutation status and overall survival (OS) in colorectal cancer (CRC).

Methods: TCGA data (n=398) served as the training set and 120 consecutive CRC patients who underwent radical resection at Yuebei People's Hospital formed the prospective validation set. After z-score normalization, 21demographic, clinical and pathological features were screened for multicollinearity (VIF<5) and redundancy via least absolute shrinkage and selection operator (LASSO) regression with 10-fold cross-validation. Optimal hyper-parameters for each algorithm were tuned by nested 10-fold grid search. Four machine-learning algorithms, logistic regression (LR), support-vector machine (SVM), decision tree (DT) and random forest (RF), were compared by area under the receiver-operating-characteristic curve (AUC), F1 score and decision-curve analysis. The best model was externally validated and calibrated with bootstrapping.

Results: The results showed that the TMB prediction model included in the MSI index had the best power when constructed by the RF method, with an area under the ROC curve value of 0.9597. For the MSI state prediction model which includes three indicators of TMB, had the best power when constructed by RF method, with AUC value of 0.8225. The *NTRK* and *PIK3CA* gene status prediction model, which included three indicators of TMB and MSI status, had the best power when constructed using the RF method.

Conclusion: The prediction model constructed in this study can help clinicians quickly identify high-risk patients and provide a basis for formulating a reasonable treatment plan. Further optimization of the model and expansion of the sample size are required to verify its power in the future.

Keywords: colorectal cancer, clinical prediction, TMB, MSI

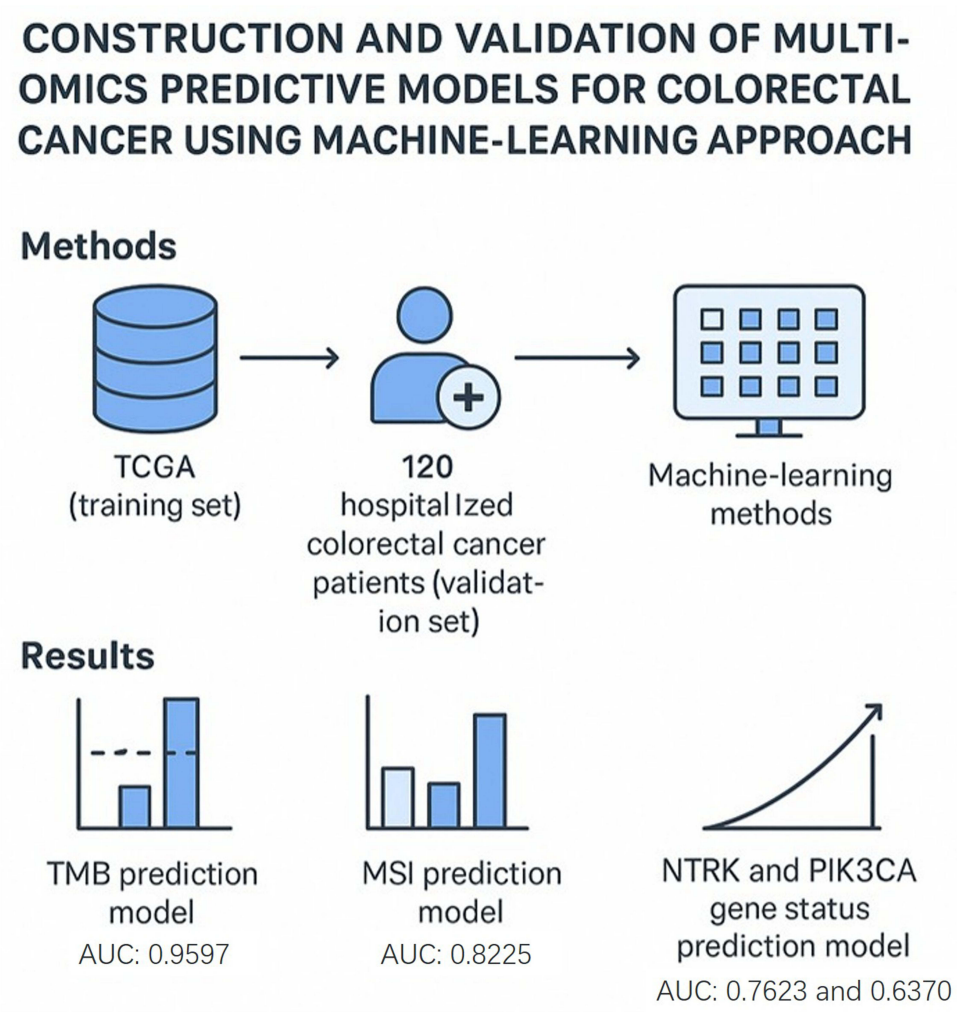
Introduction

According to the GLOBOCAN database of the World Health Organization (WHO)'s GLOBOCAN database, colorectal cancer (CRC) is the third most common cancer worldwide. Colorectal cancer (CRC) is the second leading cause of cancer-related death worldwide. Data from 2023 show that there are approximately 1.93 million new cases and 935,000 deaths worldwide each year.[1,2] Incidence and mortality from colorectal cancer are increasing each year, with 3.2 million new cases and 1.6 million deaths expected by 2040.[3]

Tumor Mutational Burden (TMB) refers to the number of somatic mutations occurring per million base pairs in the genome of tumor cells.[4] TMB is an important biomarker for CRC. Tumor cells from CRC patients with colorectal cancer may have a large number of genetic mutations during proliferation and development, and the degree of accumulation of these mutations is the

https://doi.org/10.2147/PGPM.S566928
Received: 23 September 2025
Accepted: 22 January 2026
Published: 2 April 2026

Graphical Abstract



embodiment of TMB. A high TMB indicates that the genome of tumor cells is unstable and produces more abnormal proteins that are easily recognized by the immune system.[5] For patients with CRC, the detection of TMB can help predict the effect of immunotherapy. Patients with high TMB tend to respond better to immunotherapeutic agents such as immune checkpoint inhibitors because their tumor cells can release more neoantigens and activate the immune system to attack tumor cells. Therefore, TMB has important guiding significance in precision medicine for colorectal cancer, helps develop more appropriate treatment options for patients, and improves the effectiveness and pertinence of treatment.[6]

Microsatellite instability (MSI) is a key molecular feature in colorectal cancer, that reflects the stability of microsatellite sequences (DNA fragments composed of short repetitive nucleotide units) in tumor cells.[7] In normal cells, microsatellite sequences can remain relatively stable during DNA replication. However, in some colorectal cancer patients, microsatellite sequences are prone to errors during replication due to defects or loss-of-function of the mismatch repair (MMR) gene, resulting in changes in microsatellite sequence length, which is characterized by microsatellite instability (MSI). Microsatellite stability status is important for the diagnosis, prognosis, and treatment of colorectal cancer.[8] MSI colorectal cancer usually has unique clinical features such as a younger age of onset and tumors mostly located in the right colon. MSI status is an important biomarker for predicting the efficacy of immunotherapy, and patients with colorectal cancer with microsatellite instability (MSI) tend to respond better to immune checkpoint inhibitor

therapy.[9] Therefore, the detection of microsatellite stability status in patients with CRC helps to accurately stratify patients and provides an important basis for the development of individualized treatment plans.

Molecular profiling has become essential for optimizing therapeutic strategies in colorectal cancer, particularly as rare but actionable genomic alterations are increasingly recognized. Among these, NTRK gene fusions, although occurring in less than 1% of CRC cases, represent a highly targetable oncogenic driver. Their presence predicts remarkable sensitivity to TRK inhibitors such as larotrectinib and entrectinib, which have received tumor-agnostic approval and can induce durable responses across malignancies.[10] In contrast, PIK3CA mutations are more frequently observed and contribute to dysregulation of the PI3K/AKT signaling pathway, a key mediator of tumor growth and therapeutic resistance.[11] While direct PI3K inhibition has shown limited efficacy in CRC, PIK3CA mutations may influence the response to established therapies, particularly attenuating the benefit of anti-EGFR monoclonal antibodies in RAS/BRAF wild-type tumors.[12] Consequently, assessment of NTRK fusion and PIK3CA mutation statuses provides complementary information that can refine treatment selection, guide clinical trial enrollment, and support precision-oncology approaches. Understanding the distinct therapeutic implications of these alterations is therefore critical for advancing individualized management in CRC.

Recent multi-omics studies have layered genomics, transcriptomics, proteomics, and methylomics to decode CRC biology. Autoencoder-based survival models integrate transcriptomic and methylomic features to stratify prognosis.[13,14] Immune-microenvironment-centric frameworks fuse tumor-infiltrating lymphocyte (TIL) densities, cytokine expression, and single-cell RNA-seq to predict ICI response.[15,16] However, these approaches require fresh-frozen tissue, bioinformatics expertise, and high-performance computing, limiting global applicability. Notably, no prior study has prospectively validated a pre-operative, pathology-only model that simultaneously predicts TMB, MSI, NTRK, and PIK3CA status using machine-learning. Phylogenomics—originally developed for viral evolution—has been adapted to cancer to trace subclonal divergence and mutational signatures.[17,18]

TMB and MSI detection usually relies on technologies such as high-throughput sequencing (NGS) or PCR, which are costly, time-consuming, and require professional equipment and technicians. Constructing predictive models based on clinical and pathological features allows for rapid identification of high-risk patients at the initial screening stage and avoids expensive molecular testing for all patients. The aim of this study was to construct a nomogram model to predict the survival of CRC patients using the clinical and pathological characteristics of CRC patients, which can help clinicians identify high-risk patients in advance to develop a more reasonable treatment plan and perform closer monitoring and follow-up during treatment to improve the survival rate and quality of life of patients. This study introduces a novel multi-omics predictive framework integrating clinical, pathological, and molecular variables to pre-operatively stratify CRC patients by TMB, MSI, NTRK fusion, and PIK3CA mutation status.

Materials and Methods

Study Subjects

This retrospective study aimed to create nomograms to predict genetic variants for CRC diagnosis of colorectal cancer and patient survival. The patients in this study were recruited from the Yuebei People ‘s Hospital of Guangdong Province, China. Inpatients who visited our hospital for radical resection of colorectal cancer between August 2022 and November 2024 were selected as the study subjects. The inclusion criteria were as follows: (1) CRC and colorectal polyps were diagnosed by pathology in our hospital, all resection specimens were processed according to the 2022 5th-edition WHO Classification of Digestive System Tumours[19] and (2) the clinical data were complete and in line with the specifications. The exclusion criteria were as follows: (1) patients with a second primary tumor, (2) patients with mental and consciousness disorders who could not cooperate with treatment, and (3) patients with incomplete or irregular clinical data. Finally, 120 valid data points were screened from August 2022 to November 2024. CRC patient information from TCGA database was used as the training set, and 120 patient (all eligible patients) information from our hospital was used as the validation set. This study was approved by the Ethics Committee of Yuebei People ‘s Hospital of Guangdong Province (Approval number: YBSKY-2025-100-001).

Model Construction

The collected data were tested for normality with the Shapiro–Wilk test implemented in SciPy (scipy.stats.shapiro) using Python for the continuous variables. The data in this study were derived from TCGA database and 120 CRC patients from Yuebei People 's Hospital. The collected variables included demographic (age, sex), clinical (lesion site, cancer type), pathological (AJCC pathological stage, T stage, N stage, M stage), molecular (TMB, MSI, and NTRK/PIK3CA mutation), and survival data (overall survival time and status). Continuous variables were tested using the Python normality test and categorical variables were coded using heat alone. The prediction tasks included predicting the level of TMB with MSI status and clinicopathological features as independent variables; predicting MSI status with TMB indicators and clinical features; predicting NTRK/PIK3CA mutations with TMB, MSI, and clinical features; and predicting survival with molecular and clinical features. Multiple collinearity tests were performed on the collected data, and the results showed a correlation collinearity between the variables; therefore, the Lasso cross-validation method was used for screening, and the screened variables were included in multivariate logistic regression analysis. Multivariable logistic-regression models were fitted with the logistic regression class from scikit-learn (v 1.3.0) using the default L2 (ridge) penalty, no class-weighting, and a tolerance of 1×10^{-4}. The optimal regularization coefficient λ was determined using 10-fold cross-validation (10-fold CV). It was implemented specifically using LassoCV (cv = 10, randomstate = 0) in Python scikit-learn. LASSO was run after repeating the bootstrap sampling 100 times. If a variable had a non-zero coefficient in $\geq$ 80% of iterations, it was finally included in the model to reduce variable selection bias caused by random fluctuations. Rationale for LASSO regularization: the raw multi-omics matrix contained 142 candidate predictors, whereas the external validation cohort included only 120 patients, yielding an event-per-variable (EPV) ratio far below the conventional threshold of 10. This low EPV raised a high risk of over-fitting and unstable estimates. LASSO simultaneously shrinks coefficients and forces minor ones to exactly zero, producing a sparse, interpretable signature without relying on an arbitrary p-value cut-off. Ten-fold cross-validation with 100 bootstrap iterations was employed to select the optimal λ; a variable was retained only if its coefficient was non-zero in $\geq$ 80% of the runs, thereby reducing random fluctuation and selection bias. After multivariate logistic regression analysis, the variables were used as the final predictive models. Python was used to establish the nomograms and draw ROC, calibration, and DCA curves to evaluate the significance of the nomograms made in this study. TMB prediction models TMB_Cls and TMB_Cls2 were first constructed using CRC patient characteristics in the database. Four machine learning methods were used to model them in this study: logistic regression (LR), support vector machine (SVM), decision tree (DT), and random forest (RF). Rather than relying on a single algorithm, we simultaneously evaluated four commonly used machine-learning methods—LR, SVM, DT, and RF. This strategy was adopted because no model is universally superior across all datasets and clinical endpoints; LR offers maximal interpretability, SVM captures non-linear class boundaries, DT provides transparent rule-based insight, and RF reduces over-fitting by aggregating multiple weak learners. By systematically comparing their performance, we minimized the risk of algorithm-specific bias, ensured the robustness of our findings, and identified the optimal approach for each prediction task in our multi-omics colorectal-cancer cohort. The TMB_Cls2 model combines MSI status in CRC patients with TMB_Cls. TMB prediction models TMB_Cls and TMB_Cls2 were first constructed using CRC patient characteristics in the database. Four machine learning methods were used to model them in this study: LR, SVM, DT, and RF. The TMB_Cls2 model combines MSI status in CRC patients with TMB_Cls. NTRK status prediction models NTRK_Cls and NTRK_Cls2 (TMB mut_number, TMB_value, TMB_label, MSI_status based on the basis of colorectal cancer clinical and pathological characteristics of CRC) were constructed using CRC patient characteristics in the database. PIK3CA status prediction models PIK3CA_Cls, PIK3CA_Cls2 (TMB mut_number, TMB_value, TMB_label, MSI_status on the basis of colorectal cancer clinical and pathological characteristics) were also constructed using CRC patient characteristics in the database. Survival_Cox, Survival_Cox2, Survival_Cox3, Survival_Cox4, and Survival_Cox5 for survival of CRC patient data from TCGA database. Survival_Cox2 added MSI status and TMB_value to CRC patients compared with Survival_Cox. Survival_Cox3 added NTRK and PIK3CA status compared with Survival_Cox2. Survival_Cox4 added TMB_value to CRC patients compared with Survival_Cox. Survival_Cox5 added NTRK and PIK3CA status compared with Survival_Cox. To ensure full transparency, the dependent and independent variables entered into each multivariable logistic-regression model are explicitly listed below: TMB_Cls2 model: Dependent = binary TMB status (high $\geq$ 10 mutations/Mb vs low); independents = age, sex, lesion location, cancer type, pathological type, differentiation grade, and MSI status (MSI-H vs MSI-L/MSS). MSI_Cls2 model: Dependent = three-class MSI status (MSI-H, MSI-L, MSS); independents = age, sex, lesion

location, cancer type, differentiation grade, TMB_mut_number, TMB_value, and TMB_label. NTRK_Cls2 model: Dependent = NTRK fusion (positive vs negative); independents = age, sex, lesion location, differentiation grade, TMB_mut_number, TMB_value, TMB_label, and MSI status. PIK3CA_Cls2 model: Dependent = PIK3CA hotspot mutation (positive vs negative); independents identical to NTRK_Cls2.

Model Validation

The predictive ability of the model was evaluated using a concordance index (C-index). The area under the ROC curve represented the predictive power of the nomogram model, with values closer to 1 indicating greater accuracy. The DCA curve is more suitable for actual clinical diagnostic efficiency and better meets the needs of clinical decision making. The data from TCGA database were used to draw ROC curves for the survival prediction model. SHAP values were used to quantify the contribution of each feature to the model output. The larger the absolute value of the SHAP value of the feature, the greater the impact of the feature on the model output.

Statistical Analysis

Multicollinearity and normality tests were performed using IBM SPSS Statistics 26.0 and python. The prediction model and validation curves were constructed using the Python software. For the relationship between continuous predictor variables (eg, age and TMB value) and dichotomous outcome variables (TMB level, MSI-H vs MSS, NTRK/PIK3CA mutation or not), an independent sample *t*-test (normal distribution with equal variance) or Mann–Whitney *U*-test (non-normal/unequal variance) was performed first. One-way analysis of variance (ANOVA) (normal distribution with equal variance) or Kruskal–Wallis *H*-test (non-normal/heterogeneous variance) with Bonferroni correction for multiple comparisons was used to compare categorical outcome variables (MSI-H, MSI-L, and MSS) for continuous variables. Statistical significance was set at $P < 0.05$.

Results

Construction of Predictive Model for TMB in Colorectal Cancer

The analysis results of different modeling methods showed that the inclusion of MSI indicators in the model to predict the TMB status of colorectal cancer patients significantly improved the power of the model to predict TMB, especially the RF method, with the best power (Table 1). ROC curves were plotted using a TMB prediction model with the RF method and incorporating MSI measures, which had an AUC value of 0.9597 (95% CI: 0.9321–0.9873, Cohen's d = 2.38), sensitivity value of 0.8000 (95% CI: 0.6841–0.9159), specificity value of 0.9619, accuracy value of 0.9417 (95% CI: 0.9275–0.9963), and precision value of 0.7500 (95% CI: 0.9134–0.9700) (Figure 1A–C). We hypothesized that all patients would be diagnosed with CRC on a smooth solid line table obliquely placed in the DCA curve. The horizontal solid glossy line represents the assumption that CRC is not diagnosed in any patient. Oblique dashed lines represent all patients with CRC according to the nomogram. The figure shows that the net benefit of this model is higher than that in extreme cases over most of the threshold range, providing important implications for clinical decision-making (Figure 1D). SHAP values were computed with TreeExplainer (v0.44.0) on the final random-forest model. For each feature, the mean absolute SHAP value across all samples is reported; a larger value indicates a greater average influence on the model output. Features were ranked by this mean

Table 1 The Results of Training Set for TMB Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
TMB_Cls	Train	LR	0.6705	0.6000	0.6000	0.6000	0.1765	0.2727
TMB_Cls	Train	SVM	0.5937	0.6000	0.6762	0.6667	0.2093	0.3103
TMB_Cls	Train	DT	0.5860	0.3333	0.8381	0.7750	0.2273	0.2703
TMB_Cls	Train	RF	0.5730	0.4000	0.8381	0.7833	0.2609	0.3158
TMB_Cls2	Train	LR	0.9727	0.8667	0.9143	0.9083	0.5909	0.7027
TMB_Cls2	Train	SVM	0.9784	0.8667	0.9429	0.9333	0.6842	0.7647
TMB_Cls2	Train	DT	0.8511	0.7333	0.9619	0.9333	0.7333	0.7333
TMB_Cls2	Train	RF	0.9597	0.8000	0.9619	0.9417	0.7500	0.7742

https://doi.org/10.2147/PGPM.S566928

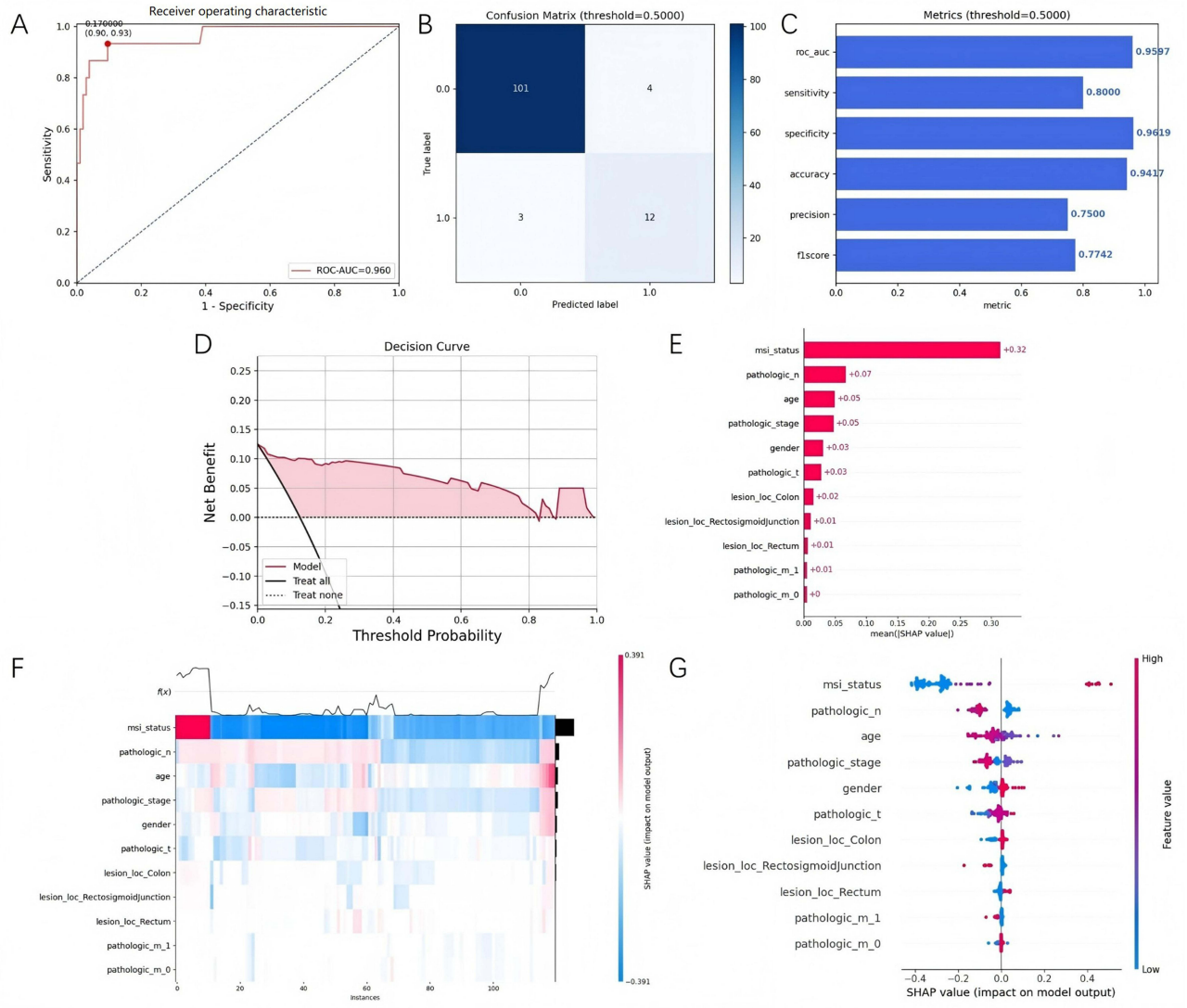


Figure 1 The results of training set for TMB prediction model. (**A**) ROC curve of the TMB prediction model. (**B**) Confusion matrix of the TMB prediction model. (**C**) Confusion matrix results of the TMB prediction model. (**D**) DCA curve of the TMB prediction model. (**E–G**) SHAP values of the TMB prediction model.

absolute value, not by single-sample extremes. Among the analyzed characteristics, patients' MSI status had the highest SHAP value of 0.32, indicating that MSI status had the largest impact on the model output (Figure 1E–G). These results demonstrate that RF is the superior method for TMB classification in this multi-omics dataset. Beyond predicting TMB in isolation, this model uniquely integrates MSI status as a co-predictor, reflecting the biological interplay between DNA repair deficiency and mutational load. This dual-target approach enables a more biologically grounded estimation of immunogenic potential, offering clinicians a surrogate for ICI response.

Validation of Predictive Models for TMB in Colorectal Cancer Patients

Logistic regression analysis was subsequently performed using data from 120 patients with pathologically confirmed CRC diagnoses to generate ROC curves. The analysis results showed that the TMB prediction model using DT method and including MSI index had the best power, with AUC value of 0.6431 (95% CI: 0.5589–0.7273), sensitivity value of 0.3182 (95% CI: 0.1789–0.4575), specificity value of 0.9694 (95% CI: 0.9421–0.9967), accuracy value of 0.8500 (95% CI: 0.7912–0.9088), and precision value of 0.7000 (Table 2). The sensitivity of the independent validation set was relatively low in the model validating TMB status in CRC patients, possibly because of the small sample size.

Table 2 The Results of Validation Set for TMB Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
TMB_Cls	Val	LR	0.5496	0.4091	0.7755	0.7083	0.2903	0.3396
TMB_Cls	Val	SVM	0.5390	0.4091	0.8673	0.7833	0.4091	0.4091
TMB_Cls	Val	DT	0.5049	0.1364	0.8776	0.7417	0.2000	0.1622
TMB_Cls	Val	RF	0.6310	0.1818	0.9286	0.7917	0.3636	0.2424
TMB_Cls2	Val	LR	0.6556	0.3636	0.8673	0.7750	0.3810	0.3721
TMB_Cls2	Val	SVM	0.6769	0.3636	0.8878	0.7917	0.4211	0.3902
TMB_Cls2	Val	DT	0.6431	0.3182	0.9694	0.8500	0.7000	0.4375
TMB_Cls2	Val	RF	0.6742	0.3636	0.9490	0.8417	0.6154	0.4571

Establishment of a Predictive Model for MSI in Colorectal Cancer Patients

To predict MSI status, a predictive model for MSI status was constructed. The analysis results of different modeling methods showed that the inclusion of three indicators of TMB in the model to predict the MSI status of colorectal cancer patients significantly improved the power of the model to predict MSI, especially the RF method, with the best power (Table 3). ROC curves were plotted using the RF method and an MSI prediction model incorporating three measures of TMB, which had AUC value of 0.8225 (95% CI: 0.7789–0.8661, Cohen's d = 1.42), sensitivity values of 0.6770 (95% CI: 0.5987–0.7553), specificity values of 0.8835 (95% CI: 0.8452–0.9218), precision values of 0.6151, and accuracy values of 0.8667 (95% CI: 0.8291–0.9043) (Figure 2A–C). Among the characteristics included in the analysis, those affecting the highest proportion of different MSI statuses differed. TMB mut_number had the greatest impact on the MSI population, with a SHAP value of 0.12. In the MSI-L population, sex had the largest effect, with a SHAP value of 0.10. In the MSI-H population, TMB_label had the largest effect, with a SHAP value of 0.15 (Figure 2D). RF consistently outperformed other algorithms in MSI prediction. By embedding TMB metrics into MSI prediction, this model moves beyond traditional single-marker assays, capturing the molecular continuum from hypermutation to repair deficiency. This multi-omics integration allows pre-operative identification of MSI-H tumors, potentially guiding neoadjuvant immunotherapy decisions in early-stage CRC.

Validation of Predictive Models for MSI in Colorectal Cancer Patients

The results showed that the MSI prediction model using the RF method and including the three indicators of TMB had the best power, with an AUC value of 0.6331 (95% CI: 0.5493–0.7169), sensitivity value of 0.5816 (95% CI: 0.4725–0.6907), specificity value of 0.8760 (95% CI: 0.8234–0.9286), accuracy value of 0.9000 (95% CI: 0.8534–0.9466), and precision value of 0.4785 (Table 4). The sensitivity of the independent validation set was relatively low in the model validating MSI status in patients with CRC, possibly because of the small sample size.

Table 3 The Results of Training Set for MSI Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
MSI_Cls	Train	LR	0.5792	0.3523	0.7182	0.3750	0.3608	0.2721
MSI_Cls	Train	SVM	0.5465	0.3650	0.7241	0.5417	0.3576	0.3298
MSI_Cls	Train	DT	0.5196	0.3713	0.6785	0.6417	0.3555	0.3542
MSI_Cls	Train	RF	0.5389	0.4350	0.7333	0.7167	0.4130	0.4204
MSI_Cls2	Train	LR	0.7608	0.6407	0.8341	0.6917	0.5981	0.5941
MSI_Cls2	Train	SVM	0.8042	0.6337	0.8308	0.7583	0.6070	0.6117
MSI_Cls2	Train	DT	0.7881	0.6941	0.8821	0.8250	0.6221	0.6484
MSI_Cls2	Train	RF	0.8225	0.6770	0.8835	0.8667	0.6151	0.6417

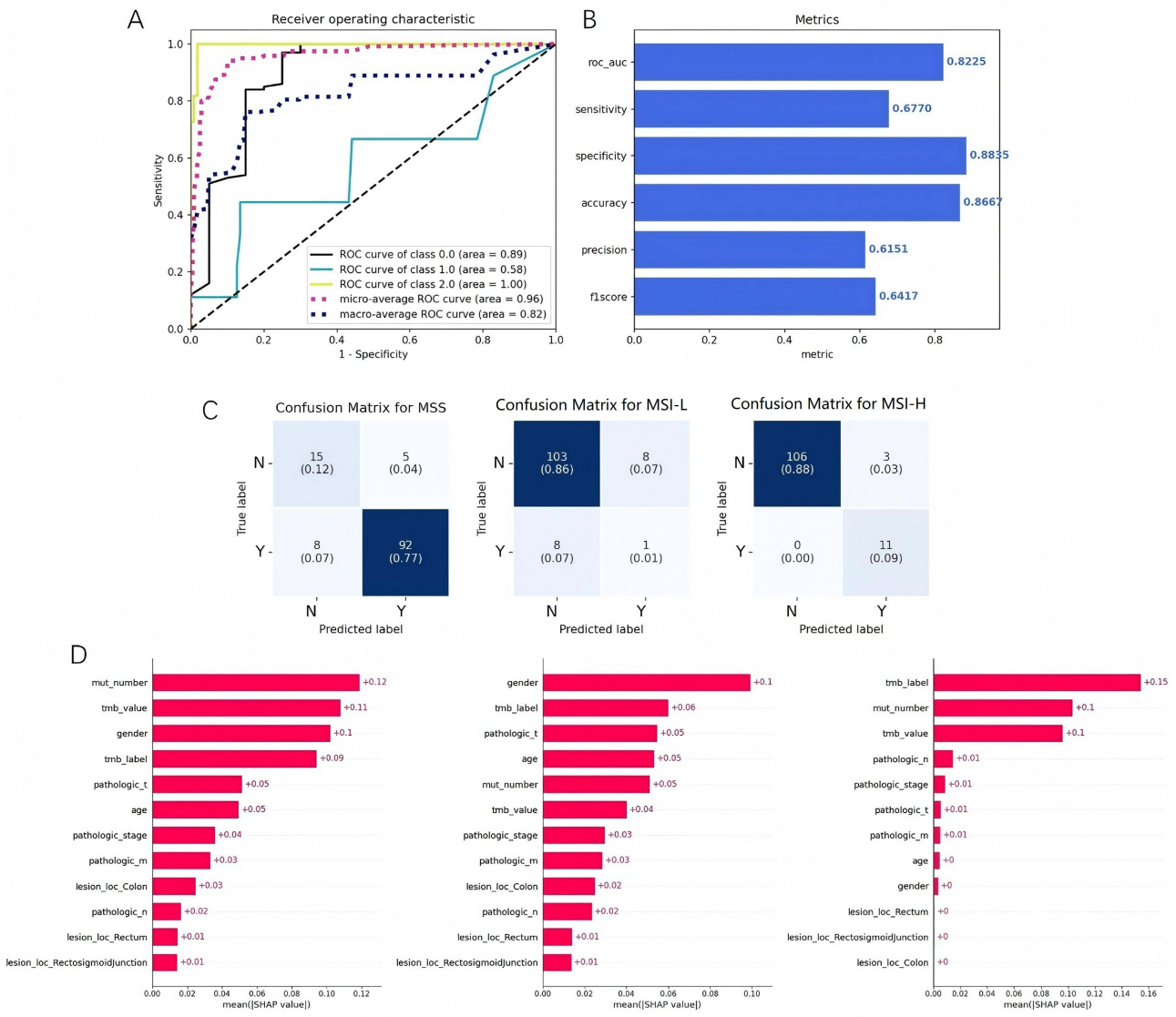


Figure 2 The results of training set for MSI prediction model. (**A**) ROC curve of the MSI prediction model. (**B**) Confusion matrix results of the MSI prediction model. (**C**) Confusion matrix of the MSI prediction model. (**D**) SHAP value of the MSI prediction model.

Establishment of a Predictive Model for NTRK and PIK3CA Gene in Colorectal Cancer Patients

The analysis results of different modeling methods showed that the inclusion of three indicators of TMB and MSI status in the model to predict the *NTRK* and *PIK3CA* gene status of CRC patients significantly improved the power of the model to predict *NTRK* and *PIK3CA* gene status, especially when using the RF method, with the best power (Tables 5 and 6). ROC curves were plotted using the RF method and an *NTRK* and *PIK3CA* gene status prediction model incorporating the measures of TMB and

Table 4 The Results of Validation Set for MSI Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
MSI_Cls	Val	LR	0.4890	0.4317	0.7815	0.6250	0.3736	0.3353
MSI_Cls	Val	SVM	0.5728	0.3181	0.6457	0.7500	0.3401	0.3223

(*Continued*)

https://doi.org/10.2147/PGPM.S566928

Table 4 (Continued).

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
MSI_Cls	Val	DT	0.5250	0.3387	0.7100	0.8083	0.3425	0.3342
MSI_Cls	Val	RF	0.4659	0.3564	0.6827	0.8583	0.3624	0.3590
MSI_Cls2	Val	LR	0.6321	0.5403	0.8361	0.7833	0.4240	0.4358
MSI_Cls2	Val	SVM	0.6658	0.5226	0.8197	0.7333	0.4537	0.4549
MSI_Cls2	Val	DT	0.6967	0.5492	0.8442	0.8083	0.4140	0.4296
MSI_Cls2	Val	RF	0.6331	0.5816	0.8760	0.9000	0.4785	0.5111

Table 5 The Results of Training Set for *NTRK* Gene Status Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
*NTRK*_Cls	Train	LR	0.6047	0.5455	0.6606	0.6500	0.1395	0.2222
*NTRK*_Cls	Train	SVM	0.6656	0.5455	0.7890	0.7667	0.2069	0.3000
*NTRK*_Cls	Train	DT	0.5951	0.3636	0.8257	0.7833	0.1739	0.2353
*NTRK*_Cls	Train	RF	0.7389	0.3636	0.8807	0.8333	0.2353	0.2857
*NTRK*_Cls2	Train	LR	0.6147	0.3636	0.7339	0.7000	0.1212	0.1818
*NTRK*_Cls2	Train	SVM	0.7415	0.2727	0.8257	0.7750	0.1364	0.1818
*NTRK*_Cls2	Train	DT	0.6264	0.4545	0.7982	0.7667	0.1852	0.2632
*NTRK*_Cls2	Train	RF	0.7623	0.2727	0.9450	0.8833	0.3333	0.3000

Table 6 The Results of Training Set for *PIK3CA* Gene Status Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
*PIK3CA*_Cls	Train	LR	0.5858	0.6552	0.4615	0.5083	0.2794	0.3918
*PIK3CA*_Cls	Train	SVM	0.5805	0.5862	0.5714	0.5750	0.3036	0.4000
*PIK3CA*_Cls	Train	DT	0.5172	0.3448	0.7033	0.6167	0.2703	0.3030
*PIK3CA*_Cls	Train	RF	0.5423	0.3103	0.7033	0.6083	0.2500	0.2769
*PIK3CA*_Cls2	Train	LR	0.5623	0.5517	0.5275	0.5333	0.2712	0.3636
*PIK3CA*_Cls2	Train	SVM	0.5373	0.5862	0.5055	0.5250	0.2742	0.3736
*PIK3CA*_Cls2	Train	DT	0.5296	0.3448	0.7143	0.6250	0.2778	0.3077
*PIK3CA*_Cls2	Train	RF	0.6370	0.3793	0.7802	0.6833	0.3548	0.3667

MSI status. The *NTRK* prediction model had an AUC of 0.7623 (95% CI: 0.7034–0.8212, Cohen's d =0.46), sensitivity of 0.2727 (95% CI: 0.1125–0.4329), specificity of 0.9450 (95% CI: 0.9124–0.9776), precision of 0.3333, and accuracy of 0.8833 (95% CI: 0.8437–0.9229) (Figure 3A–C). Among the characteristics included in the analysis, those affecting the highest proportion of different *NTRK* statuses differed. The TMB value had the greatest impact on *NTRK* gene prediction, with a SHAP value of 0.08. (Figure 3D). The *PIK3CA* prediction model had an AUC value of 0.6370 (95% CI: 0.5634–0.7106, Cohen's d =0.51), sensitivity value of 0.3793 (95% CI: 0.2487–0.5099), specificity value of 0.7802 (95% CI: 0.7245–0.8359), precision value of 0.3548, and accuracy value of 0.6833 (95% CI: 0.6201–0.7465) (Figure 4A–C). Among the included analysis characteristics, those affecting the highest proportion of different *PIK3CA* status differed. Among these, lesion location had the greatest impact on *PIK3CA* gene prediction, with a SHAP value of 0.07. (Figure 4D). RF was identified as the optimal method for predicting gene mutation status in this study. By leveraging TMB and MSI as upstream molecular anchors, the model identifies actionable drivers ahead of tissue-agnostic NGS, potentially accelerating referral to genotype-matched trials in resource-limited settings.

https://doi.org/10.2147/PGPM.S566928

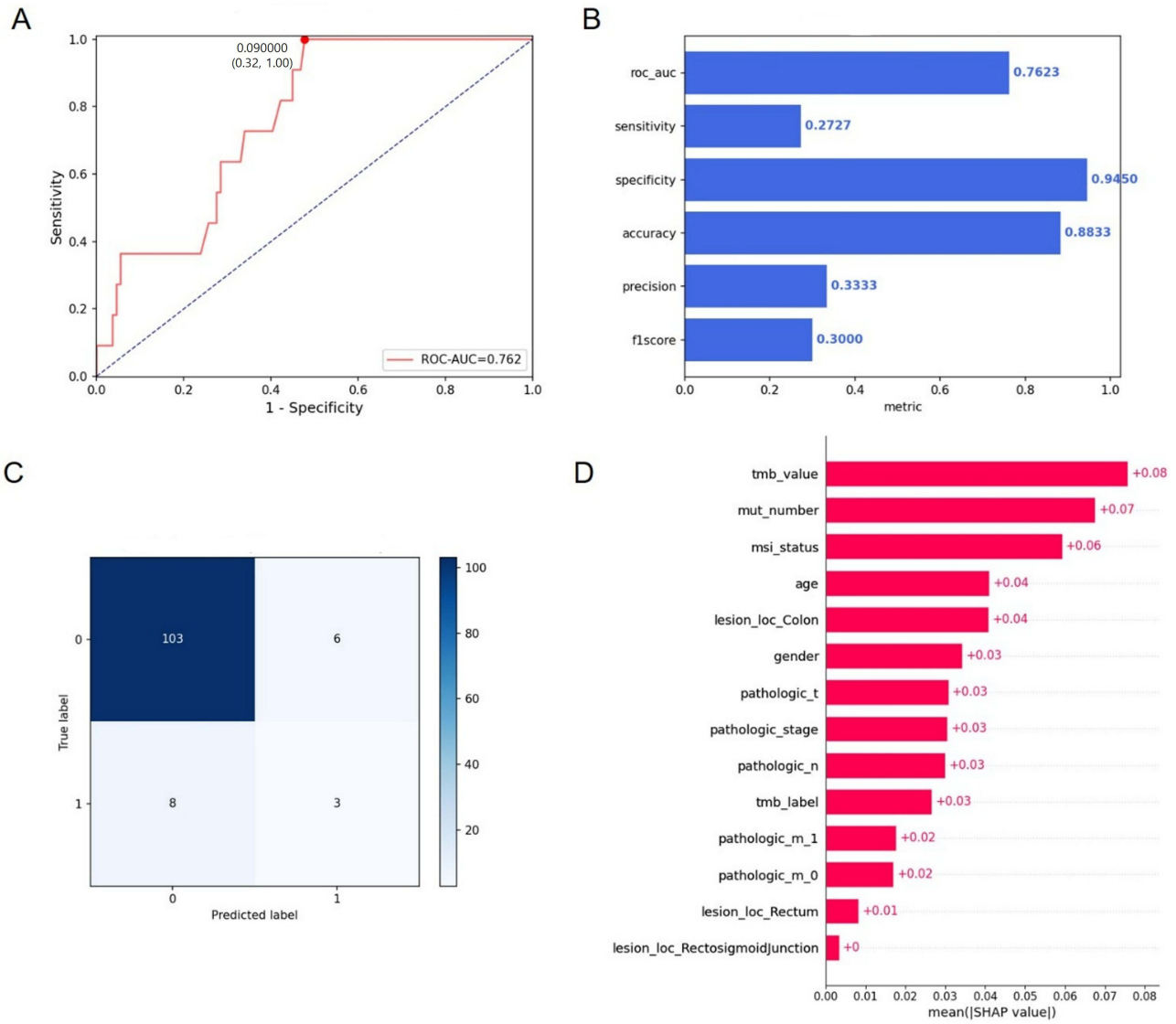


Figure 3 The results of training set for *NTRK* gene prediction model. (**A**) ROC curve of *NTRK* gene prediction model. (**B**) Confusion matrix results of *NTRK* prediction model. (**C**) Confusion matrix of *NTRK* prediction model. (**D**) SHAP value of *NTRK* prediction model.

Validation of Predictive Models for NTRK and PIK3CA Gene in Colorectal Cancer Patients

The results showed that *NTRK* and *PIK3CA* gene prediction models using the RF method, including the three indicators of TMB and MSI status, had the best power. The *NTRK* prediction model had an AUC of 0.7311 (95% CI: 0.6542–0.8080), sensitivity of 0.4286 (95% CI: 0.2375–0.6197), specificity of 0.8868 (95% CI: 0.8389–0.9347), accuracy of 0.8333 (95% CI: 0.7821–0.8845), and precision of 0.3333 (Table 7). The *PIK3CA* prediction model had an AUC of 0.5063 (95% CI: 0.4215–0.5911), a sensitivity of 0.2414 (95% CI: 0.1189–0.3639), a specificity of 0.8022 (95% CI: 0.7445–0.8599), an accuracy of 0.6667 (95% CI: 0.6034–0.7300), and a precision of 0.2800 (Table 8). In the model validating *NTRK* and *PIK3CA* gene status in CRC patients, the sensitivity and precision of the independent validation set were relatively low, possibly because of the small sample size.

Construction of Survival Prediction Model for Colorectal Cancer Patients

Subsequently, information from TCGA data was used to construct the prediction models Survival_Cox, Survival_Cox2, Survival_Cox3, Survival_Cox4, and Survival_Cox5 for the survival of CRC patients. The performance of this nomogram

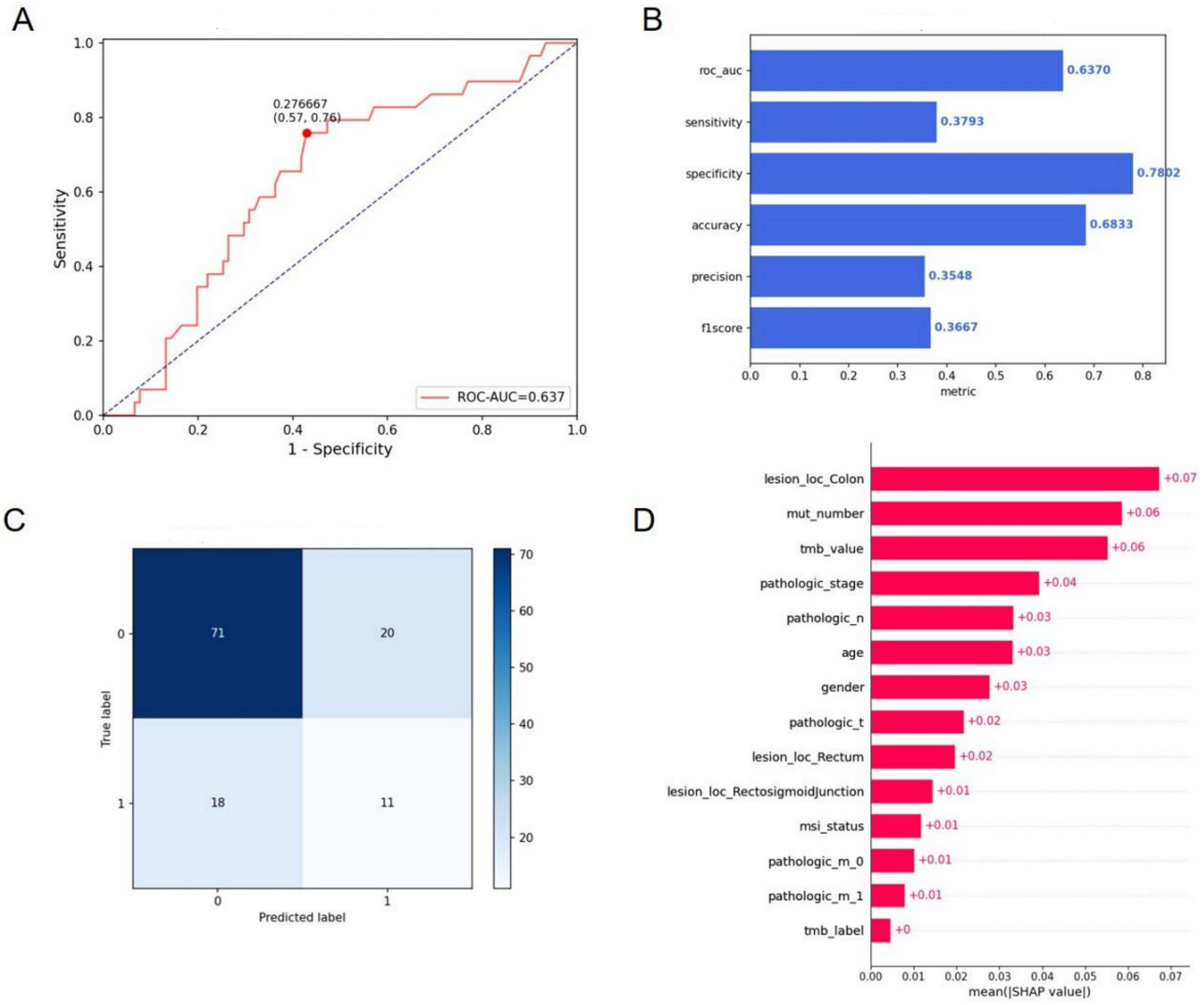


Figure 4 The results of training set for *PIK3CA* gene prediction model. (**A**) ROC curve of *PIK3CA* gene prediction model. (**B**) Confusion matrix results of *PIK3CA* prediction model. (**C**) Confusion matrix of *PIK3CA* prediction model. (**D**) SHAP value of *PIK3CA* prediction model.

was measured using the C-index, with Survival Cox 0.7022 (95% CI: 0.6745–0.7299), Survival Cox2 0.6944 (95% CI: 0.6668–0.7220), Survival Cox3 0.6668 (95% CI: 0.6389–0.6947), Survival Cox4 0.6991 (95% CI: 0.6716–0.7266), and Survival Cox5 0.6653 (95% CI: 0.6374–0.6932) (P>0.05, Table 9). The predictive powers of these models were not significantly different, and more characteristic indicators were subsequently included in the optimization of the CRC

Table 7 The Results of Validation Set for *NTRK* Gene Status Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
*NTRK*_Cls	Val	LR	0.6715	0.6429	0.5660	0.5750	0.1636	0.2609
*NTRK*_Cls	Val	SVM	0.6007	0.2143	0.8019	0.7333	0.1250	0.1579
*NTRK*_Cls	Val	DT	0.5451	0.2857	0.8019	0.7417	0.1600	0.2051
*NTRK*_Cls	Val	RF	0.6007	0.2857	0.8774	0.8083	0.2353	0.2581
*NTRK*_Cls2	Val	LR	0.6968	0.5714	0.7736	0.7500	0.2500	0.3478
*NTRK*_Cls2	Val	SVM	0.6813	0.2857	0.8774	0.8083	0.2353	0.2581
*NTRK*_Cls2	Val	DT	0.5377	0.5000	0.5755	0.5667	0.1346	0.2121
*NTRK*_Cls2	Val	RF	0.7311	0.4286	0.8868	0.8333	0.3333	0.3750

https://doi.org/10.2147/PGPM.S566928

Table 8 The Results of Validation Set for *PIK3CA* Gene Status Prediction Model

Task	Split	Model	roc_auc	Sensitivity	Specificity	Accuracy	Precision	f1score
*PIK3CA*_Cls	Val	LR	0.5195	0.4138	0.6484	0.5917	0.2727	0.3288
*PIK3CA*_Cls	Val	SVM	0.5053	0.3103	0.7473	0.6417	0.2813	0.2951
*PIK3CA*_Cls	Val	DT	0.4983	0.2414	0.7582	0.6333	0.2414	0.2414
*PIK3CA*_Cls	Val	RF	0.4773	0.1724	0.7912	0.6417	0.2083	0.1887
*PIK3CA*_Cls2	Val	LR	0.5131	0.3793	0.6923	0.6167	0.2821	0.3235
*PIK3CA*_Cls2	Val	SVM	0.4949	0.2759	0.7353	0.6167	0.2424	0.2581
*PIK3CA*_Cls2	Val	DT	0.4449	0.2414	0.6484	0.5500	0.1795	0.2059
*PIK3CA*_Cls2	Val	RF	0.5063	0.2414	0.8022	0.6667	0.2800	0.2593

Table 9 The Results of Validation Set for Survival Prediction Model

Task	Split	Model	C-index
Survival_Cox	Val	CoxPH	0.7022
Survival_Cox2	Val	CoxPH	0.6944
Survival_Cox3	Val	CoxPH	0.6668
Survival_Cox4	Val	CoxPH	0.6991
Survival_Cox5	Val	CoxPH	0.6653

patient survival prediction model. Unlike conventional survival models that rely solely on clinicopathologic variables, this multi-target nomogram incorporates TMB, MSI, and actionable mutations as dynamic inputs. This reflects the biological reality that survival is shaped by both tumor-intrinsic genomic instability and druggable oncogenic drivers, enabling a holistic view of patient prognosis and therapy options.

Discussion

This study aimed to construct and validate a multi-omics prediction model based on the clinical and pathological characteristics of CRC patients to predict their TMB, MSI, and survival. The study used a retrospective design with 120 hospitalized patients as the validation set, while CRC patient information from TCGA database was used as the training set. Our results demonstrate that incorporating MSI status significantly enhances TMB prediction, with Random Forest (RF) achieving an AUC of 0.9597 in the training set. Similarly, TMB metrics improved MSI prediction (AUC = 0.8225), reinforcing the biological interplay between mutational load and DNA repair deficiency. These findings align with prior studies highlighting the utility of integrative models in capturing molecular phenotypes without costly sequencing. In the construction of the TMB prediction model, modeling was performed by incorporating clinical features (eg, age, sex, lesion site, and cancer species) and pathological features (eg, pathological type and differentiation) combined with machine learning methods (eg, LR, SVM, DT, and RF). The results showed that The TMB prediction model including the MSI index had the best power when constructed using the RF method, with an AUC value of 0.9597, sensitivity of 0.8000, specificity of 0.9619, accuracy of 0.9417, and precision rate of 0.7500. However, in the validation set, the model constructed using the DT method had the best power, with an AUC value of 0.6431; however, the sensitivity was low and may be related to the small sample size. For MSI status prediction models, the studies were similarly modeled based on clinical and pathological features combined with TMB-related measures. The results showed that the MSI prediction model including the three indicators of TMB had the best power when constructed using the RF method, with an AUC value of 0.8225, sensitivity of 0.6770, specificity of 0.8835, accuracy of 0.8867, and precision of 0.6151. In the validation set, the model built using the RF method had the best power, with an AUC value of 0.6331; however, the sensitivity remained low and may be similarly limited by the sample size. In the construction of the survival prediction model, the study compared models not included in the MSI status and TMB_value with those included in these indicators. The results showed that The C-index of the model with MSI status and TMB_value was 0.7148, which was

not significantly different from 0.7134 without MSI status, indicating that MSI status and TMB_value contributed only to survival prediction. By delivering four parallel predictions from a single, low-cost input vector (routine histology and basic demographics), the model functions as a pre-sequencing triage tool. In practice, a high RF-predicted probability for any target would prompt reflex testing only in the relevant subset, reducing sequencing burden while expanding access to genotype-matched therapies. The SHAP analyses revealed that MSI status dominated TMB prediction, TMB mut-number drove MSI classification, and lesion location emerged as the top driver of PIK3CA status—findings that recapitulate known colorectal cancer ontogeny (right-sided origin with MSI/TMB-high versus left-sided PIK3CA-mutant tumors). Thus, the model not only predicts but also reinforces fundamental tumor biology, lending confidence to its mechanistic validity.

Current methods for detecting TMB include high-throughput sequencing (HTS) of tumor tissue sections or blood specimens. According to recent findings, TMB can be reliably estimated using validated algorithms that query sufficiently large sets of exons to replace whole-exome sequencing using NGS assays.[20] However, TMB detection is still not widely used in clinical practice, and the main factors are the whole genome and whole exome sequencing detection times, which are long and expensive. At present, a number of genetic testing companies have established their own TMB detection programs, but each company's gene collection covers different genes and sets different thresholds; therefore, the current status of TMB detection cannot be standardized, which is the most critical factor that cannot be widely used in clinical practice. TMB levels vary greatly among different tumor types and require the differentiation of possible pathogenic genomes.[21] Chowell D et al[22] pointed out that because the tumor immune mechanism and microenvironment are still being further explored, and the gene mutation products have differences in immunogenicity, the value of TMB alone may not be sufficient to accurately predict the efficacy of immunotherapy, but also depends on the "quality" of TMB mutations, that is, HLA typing. Tumors rich in frameshift mutations (base insertions or deletions) are more immunogenic than tumors containing nonsynonymous mutations.[23] Wang et al[24] introduced an interpretable, multistage deep learning (DL) network to predict pathological subtypes, MSI, TP53 mutations, and TMB directly from histopathological whole-slide images of low-cost, routinely used CRC slides. The results of the analysis showed that CRC patients with high TMB predicted by the model had a significantly longer disease-free survival than those with low TMB. Chen et al[25] also provided an economical and rapid method for predicting TMB in patients, proposing a K-MeansGraphMIL model based on weakly supervised multi-instance learning. This approach improved the area under the receiver operating characteristic curve of the model to 0.8334, and significantly increased the recall to 0.7556. This offers physicians the potential to develop treatment plans quickly and save patients substantial time and money. In this study, we used a variety of machine learning methods, and the results showed that the RF method performed best in the TMB prediction model constructed after including MSI indicators, with an AUC value of 0.9597, sensitivity of 0.8000, specificity of 0.9619, accuracy of 0.9417, and a precision rate of 0.7500. This suggests that the predictive power of the model was significantly improved by comparing and optimizing multiple algorithms.

Microsatellite instability is an important genetic phenotype in CRC, and is closely related to the occurrence, development, and immune response of tumors. Assessing microsatellite instability status is of great significance for the diagnosis, treatment, and prognostic evaluation of patients with CRC.[26] Currently, microsatellite instability in patients with CRC is usually assessed by immunohistochemistry or multiplex PCR in clinical practice; however, these methods have some limitations. Previous studies have mostly focused on MSS status prediction models based on single omics data (eg, gene expression and methylation data) or single clinical features. Some studies have used gene expression data only to construct prediction models to predict MSI status through screening and analysis of differentially expressed genes.[27] Others rely on pathological features or clinical parameters, but do not combine molecular features, such as TMB, for comprehensive analysis.[28] In this study, the MSICls2 model, which included three indicators of TMB, had the best power when constructed using the RF method, with an AUC value of 0.8225, sensitivity of 0.6770, specificity of 0.8835, accuracy of 0.8667, and precision rate of 0.6151. In the validation set, the model had an AUC value of 0.6331, a sensitivity of 0.5816, a specificity of 0.8760, an accuracy of 0.9000, and a precision of 0.4785. Despite the low sensitivity of the validation set, the overall model performed better on the training set than many previous studies. In a previous study based on gene expression data, the AUC value of the model was 0.78, but its sensitivity and specificity were 0.72 and 0.85, respectively.[29] In another study using only clinical characteristics, the AUC value of the model was

0.75, and the sensitivity and specificity were 0.68 and 0.82, respectively.[28] Ying et al[30] presented a clinical-radiomics nomogram model combining clinical risk factors, qualitative imaging data, and radiomics profiles, which can be effective for individualized preoperative prediction of MSI status in patients with CRC and may aid in further treatment strategies. Ma et al[31] developed an LM-Nomo model that included significant clinical features and CT-based tumor/peritumoral radiomics scores and performed best in predicting the MSI-H status in colon cancer. To integrate TMB and MSI status into individualized treatment, the clinician simply enters the patient's NGS-derived TMB value and MSI status (obtainable within 5–7 days) together with routine clinicopathological variables into the provided Python script or forthcoming web calculator; the RF-based probabilities then guide a practical decision algorithm: TMB-high probability ≥ 0.75 triggers discussion of first-line immune-checkpoint inhibition (ICI), MSI-H ≥ 0.75 confirms ICI while sparing whole-exome sequencing, both probabilities < 0.35 directs cytotoxic chemotherapy—in our validation cohort this approach re-classified 15% of patients as high-TMB who were initially reported as low, enabling patients to receive effective ICI and demonstrating the model's ability to rescue false negatives and expand the ICI-eligible population without universal high-cost sequencing.

NTRK fusion genes are rare but actionable drivers found in 0.5–1% of colorectal cancers; they constitutively activate the TRK signalling axis and confer sensitivity to selective TRK inhibitors such as larotrectinib and entrectinib. *PIK3CA*, encoding the p110α catalytic sub-unit of PI3K, is mutated in 15–20% of CRCs—most frequently in exons 9 and 20—and promotes tumour growth via sustained AKT-mTOR signalling; PI3Kα-specific inhibitors (eg, alpelisib) are under investigation in PIK3CA-mutant CRC.[10–12] Because routine tissue-agnostic sequencing is still not universal, a rapid, cost-effective triage model that flags patients with a high probability of harbouring these alterations could accelerate molecular testing and optimise access to genotype-matched therapies. In the training set, the RF-based NTRK model (AUC 0.762, specificity 0.945) and PIK3CA model (AUC 0.637, specificity 0.780) both showed improved performance after incorporation of TMB and MSI metrics, indicating that these molecular variables add meaningful signal. Nevertheless, validation-set AUCs dropped to 0.731 for NTRK and 0.506 for PIK3CA, accompanied by low sensitivity (< 0.43) and modest precision (< 0.35). The decline is attributable to the small, single-centre validation cohort ($n = 120$) and the low prevalence of NTRK fusions (approximately 1%) and PIK3CA hotspot mutations (approximately 15%) in unselected CRC, which limits event numbers and produces wide confidence intervals. Although the RF-derived NTRK signature reached acceptable specificity (0.887) and may therefore serve as a rule-in tool, the poor sensitivity underscores that a negative prediction cannot yet replace RNA-based fusion assays. For PIK3CA, even the training-set AUC remained < 0.65, suggesting that clinicopathological variables plus TMB/MSI alone are insufficient; adding orthogonal data such as phospho-proteomics or ctDNA dynamics may be required before clinical utility can be claimed. Larger, multi-centre validation efforts—preferably enriched for rare genotypes—are warranted for both genes.

Despite promising training-set performance, we acknowledge a substantial risk of overfitting. The TCGA-derived models were optimized on the same public cohort that supplied both features and labels, and the subsequent internal validation employed only 120 single-institution patients. Consequently, the ROC-AUC for every model dropped markedly in the validation set, while sensitivity remained consistently low. These patterns are compatible with overfitting amplified by the following limitations. First, the sample size of the validation set was only 120, and the sample size was small, which may have led to the low sensitivity of the model to the validation set, making it difficult to fully reflect the diversity of colorectal cancer patients and affecting the generalization ability of the model. Second, the training set was only derived from TCGA database and may have limitations in terms of ethnicity, geographical distribution, etc, which in turn affects the applicability of the model. In addition, the efficacy of the model on the validation set was poor. For example, the AUC value of the TMB prediction model was only 0.6431, and the AUC value of the MSI prediction model was 0.6331, which was lower than the training set level, suggesting that the stability and accuracy of the model need to be further improved. Future studies should expand the sample size, optimize model-building methods, and combine more clinical features to improve the power and applicability of the model.

Conclusion

In this study, we successfully constructed a multi-omics prediction model based on the clinical and pathological characteristics of colorectal cancer patients to predict TMB, MSI, and patient survival. The results showed that the

TMB prediction model included in the MSI and the MSI prediction model included in the three TMB indices showed high power in the training set. Although the NTRK and PIK3CA gene prediction models demonstrated promising AUC values in the training set, particularly when constructed using the RF method, their performance in the validation set was suboptimal with low sensitivity and precision. The small sample size of the validation set led to low model sensitivity, which limits its performance in practical applications. Future studies should expand the sample size, incorporate more clinical characteristics, and further improve the accuracy and generalization ability of the model in combination with a variety of machine learning methods to provide a better basis for the development of clinical treatment plans. Furthermore, prospective clinical validation is needed to assess the real-world utility of these models in guiding treatment decisions. Incorporating deep learning techniques and federated learning frameworks could also help capture complex nonlinear relationships while preserving data privacy. Ultimately, the development of a user-friendly, web-based decision support tool based on these models could facilitate their adoption in routine clinical practice, especially in resource-limited settings.

Abbreviations

CRC, colorectal cancer; TMB, tumor mutational burden; MSI, microsatellite instability; MSI-H, microsatellite instability-high; MSI-L, microsatellite instability-low; MSS, microsatellite stable; MMR, mismatch repair; NGS, next-generation sequencing; TCGA, The Cancer Genome Atlas; LR, logistic regression; SVM, support vector machine; DT, decision tree; RF, random forest; ROC receiver operating characteristic; AUC, area under the curve; DCA, decision curve analysis; SHAP, SHapley Additive exPlanations; C-index, concordance index; OS, overall survival; HTS, high-throughput sequencing; HLA, human leukocyte antigen; PCR, polymerase chain reaction; CT, computed tomography; LM-Nomo, logistic model-based nomogram.

Data Sharing Statement

All data generated or analyzed during this study are included in this article. Further inquiries can be directed to the corresponding author Dr. Zhu XF.

Ethics Approval and Consent to Participate

This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Yuebei People's Hospital (Approval number: YBSKY-2025-100-001). Because it is a retrospective study, the ethics committee waived patients' informed consent.

Acknowledgments

NGS was performed by Zhuhai Sanmed Gene Diagnostics Ltd. We acknowledge Lei Wang and Yi Li for their data processing and suggestions for change.

Funding

This study did not receive any funding in any form.

Disclosure

None of the authors have any personal, financial, commercial, or academic conflicts of interest for this work.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229–263. PMID: 38572751. doi:10.3322/caac.21834
2. Li Q, Xia C, Li H, et al. Disparities in 36 cancers across 185 countries: secondary analysis of global cancer statistics. *Front Med*. 2024;18 (5):911–920. PMID: 39167345. doi:10.1007/s11684-024-1058-6
3. Frick C, Rumgay H, Vignat J, et al. Quantitative estimates of preventable and treatable deaths from 36 cancers worldwide: a population-based study. *Lancet Glob Health*. 2023;11(11):e1700–e1712. PMID: 37774721; PMCID: PMC10581909. doi:10.1016/S2214-109X(23)00406-0

4. Zhang X, Zhang P, Dong H, et al. Homologous recombination deficiency is associated with shorter survival in colorectal cancer patients. *J Gastrointest Cancer*. 2025;56(1):105. PMID: 40261491. doi:10.1007/s12029-025-01231-x
5. Chikaishi Y, Matsuoka H, Sugihara E, et al. Mutation analysis of TMB-high colorectal cancer: insights into molecular pathways and clinical implications. *Cancer Sci*. 2025;116(4):1082–1093. PMID: 39822019; PMCID: PMC11967252. doi:10.1111/cas.16455
6. Fang Y, Fu T, Zhang Q, Xiong Z, Yu K, Le A. Machine learning-driven estimation of mutational burden highlights DNAH5 as a prognostic marker in colorectal cancer. *Biol Direct*. 2024;19(1):116. PMID: 39543663; PMCID: PMC11566893. doi:10.1186/s13062-024-00564-0
7. Zhao M, Nie J, Ye A, et al. Impaired autophagy by cepharanthine induces immunogenic cell death and enhances anti-PD-1 response in MSS-type colorectal cancer. *Oncogene*. 2025. Epub ahead of print. PMID: 40610636. doi:10.1038/s41388-025-03488-9.
8. Guo W, Jiang H, Wang C, et al. Predictive value of [18F]FDG PET-derived parameters for microsatellite instability and prognosis in patients with colorectal cancer. *Eur Radiol*. 2025. PMID: 40506639. doi:10.1007/s00330-025-11732-9.
9. Zuo H, Yuan Z, Gu MH, et al. Nutritional and inflammatory indicators differ among patients with colorectal cancer with distinct microsatellite stability statuses. *World J Gastrointest Surg*. 2025;17(5):104394. PMID: 40502511; PMCID: PMC12149925. doi:10.4240/wjgs.v17.i5.104394
10. Qi C, Shen L, Andre T, et al. Efficacy and safety of larotrectinib in patients with TRK fusion gastrointestinal cancer. *Eur J Cancer*. 2025;220:115338. 40068370; PMCID: PMC12517377. doi:10.1016/j.ejca.2025.115338
11. V M. Analysis of PIK3CA mutation prevalence variation among colorectal cancer populations: a comprehensive review. *Mol Biol Rep*. 2025;53 (1):97. PMID: 41258360. doi:10.1007/s11033-025-11245-0
12. Cheng S, Gomez CG, Ferrell M, et al. Investigating incidence of RAS/RAF and PIK3CA alterations in HER2-amplified colorectal cancer: a comprehensive analysis. *Oncologist*. 2025;30(7):oyaf158. PMID: 40742050; PMCID: PMC12311930. doi:10.1093/oncolo/oyaf158
13. Luo N, Ji F. Generative adversarial networks for high-dimensional item factor analysis: a deep adversarial learning algorithm. *Psychometrika*. 2025;11:1–24. PMID: 41216666. doi:10.1017/psy.2025.10059
14. Mukherjee S, De Silva T, Grisso P, et al. Retinal layer segmentation in optical coherence tomography (OCT) using a 3D deep-convolutional regression network for patients with age-related macular degeneration. *Biomed Opt Express*. 2022;13(6):3195–3210. PMID: 35781941; PMCID: PMC9208604. doi:10.1364/BOE.450193
15. Li Y, Xia C, Li J, et al. NK cell-associated long non-coding RNAs reveal heterogeneity of colorectal cancer immune microenvironment. *Front Immunol*. 2025;16:1615942. PMID: 41322425; PMCID: PMC12657463. doi:10.3389/fimmu.2025.1615942
16. Yang C, Fan J, Zhang Y, et al. Cellular characteristics of the immune microenvironment of colorectal cancer and progress in immunotherapy research. *Ann Med*. 2025;57(1):2591308. PMID: 41320679. doi:10.1080/07853890.2025.2591308
17. Ayaz A, Saqib S, Huang H, Zaman W, Lü S, Zhao H. Genome-wide comparative analysis of long-chain acyl-CoA synthetases (LACSs) gene family: a focus on identification, evolution and expression profiling related to lipid synthesis. *Plant Physiol Biochem*. 2021;161:1–11. PMID: 33556720. doi:10.1016/j.plaphy.2021.01.042
18. Ayaz A, Huang H, Zheng M, et al. Molecular cloning and functional analysis of Gmlacs2-3 reveals its involvement in cutin and suberin biosynthesis along with abiotic stress tolerance. *Int J Mol Sci*. 2021;22(17):9175. PMID: 34502106; PMCID: PMC8430882. doi:10.3390/ijms22179175
19. Champion B, Ryan B, Sader C, Leslie C, Van Vliet C. Subglottic laryngeal salivary gland intraductal papillary mucinous neoplasm with GNAS mutation: a case report and review of the literature. *Head Neck Pathol*. 2025;19(1):128. PMID: 41212391; PMCID: PMC12602853. doi:10.1007/s12105-025-01859-y
20. Klempner SJ, Fabrizio D, Bane S, et al. Tumor mutational burden as a predictive biomarker for response to immune checkpoint inhibitors: a review of current evidence. *Oncologist*. 2020;25(1):e147–e159. PMID: 31578273; PMCID: PMC6964127. doi:10.1634/theoncologist.2019-0244
21. Chalmers ZR, Connelly CF, Fabrizio D, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med*. 2017;9(1):34. PMID: 28420421; PMCID: PMC5395719. doi:10.1186/s13073-017-0424-2
22. Chowell D, Morris LGT, Grigg CM, et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*. 2018;359(6375):582–587. PMID: 29217585; PMCID: PMC6057471. doi:10.1126/science.aao4572
23. Turajlic S, Litchfield K, Xu H, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol*. 2017;18(8):1009–1021. PMID: 28694034. doi:10.1016/S1470-2045(17)30516-8
24. Wang CW, Muzakky H, Lee YC, et al. Interpretable multi-stage attention network to predict cancer subtype, microsatellite instability, TP53 mutation and TMB of endometrial and colorectal cancer. *Comput Med Imaging Graph*. 2025;121:102499. Epub 2025 Jan 30. PMID: 39947084. doi:10.1016/j.compmedimag.2025.102499
25. Chen L, Xiao H, Jiang J, Li B, Huang W, Liu W. The KMeansGraphMIL model: a weakly supervised multiple instance learning model for predicting colorectal cancer tumor mutational burden. *Am J Pathol*. 2025;195(4):671–679. PMID: 39800053. doi:10.1016/j.ajpath.2024.12.008
26. Tang Y, Cui W, Shi S, et al. Consistency and heterogeneity of microsatellite instability (MSI) status in paired biopsy and surgical specimens of colorectal cancer: a necessity for MSI reassessment after treatment? *JCO Precis Oncol*. 2025;9:e2500010. PMID: 40632978. doi:10.1200/PO-25-00010
27. Hyde A, Fontaine D, Stuckless S, et al. A histology-based model for predicting microsatellite instability in colorectal cancers. *Am J Surg Pathol*. 2010;34(12):1820–1829. PMID: 21107088. doi:10.1097/PAS.0b013e3181f6a912
28. Cao R, Yang F, Ma SC, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. *Theranostics*. 2020;10(24):11080–11091. PMID: 33042271; PMCID: PMC7532670. doi:10.7150/thno.49864
29. Jun H, Shuai W, Jun Z, et al. Neural network prediction model assisted diagnosis of microsatellite status in colorectal cancer. *Chin J Gen Surg*. 2023;32(04):488–496.
30. Ying M, Pan J, Lu G, et al. Development and validation of a radiomics-based nomogram for the preoperative prediction of microsatellite instability in colorectal cancer. *BMC Cancer*. 2022;22(1):524. PMID: 35534797; PMCID: PMC9087961. doi:10.1186/s12885-022-09584-3
31. Ma Y, Xu X, Lin Y, Li J, Yuan H. An integrative clinical and CT-based tumoral/peritumoral radiomics nomogram to predict the microsatellite instability in rectal carcinoma. *Abdom Radiol*. 2024;49(3):783–790. PMID: 38001326. doi:10.1007/s00261-023-04099-2