



OPEN

Exploring the value of multiple preprocessors and classifiers in constructing models for predicting microsatellite instability status in colorectal cancer

Yi Ma^{1,3}, Zhihao Shi^{1,3}, Ying Wei², Feng Shi², Guochu Qin¹✉ & Zhengyang Zhou¹✉

Approximately 15% of patients with colorectal cancer (CRC) exhibit a distinct molecular phenotype known as microsatellite instability (MSI). Accurate and non-invasive prediction of MSI status is crucial for cost savings and guiding clinical treatment strategies. The retrospective study enrolled 307 CRC patients between January 2020 and October 2022. Preoperative images of computed tomography and postoperative status of MSI information were available for analysis. The stratified fivefold cross-validation was used to avoid sample bias in grouping. Feature extraction and model construction were performed as follows: first, inter-/intra-correlation coefficients and the least absolute shrinkage and selection operator algorithm were used to identify the most predictive feature subset. Subsequently, multiple discriminant models were constructed to explore and optimize the combination of six feature preprocessors (Box-Cox, Yeo-Johnson, Max-Abs, Min-Max, Z-score, and Quantile) and three classifiers (logistic regression, support vector machine, and random forest). Selecting the one with the highest average value of the area under the curve (AUC) in the test set as the radiomics model, and the clinical screening model and combined model were also established using the same processing steps as the radiomics model. Finally, the performances of the three models were evaluated and analyzed using decision and correction curves. We observed that the logistic regression model based on the quantile preprocessor had the highest average AUC value in the discriminant models. Additionally, tumor location, the clinical of N stage, and hypertension were identified as independent clinical predictors of MSI status. In the test set, the clinical screening model demonstrated good predictive performance, with the average AUC of 0.762 (95% confidence interval, 0.635–0.890). Furthermore, the combined model showed excellent predictive performance (AUC, 0.958; accuracy, 0.899; sensitivity, 0.929) and favorable clinical applicability and correction effects. The logistic regression model based on the quantile preprocessor exhibited excellent performance and repeatability, which may further reduce the variability of input data and improve the model performance for predicting MSI status in CRC.

Keywords Microsatellite instability, Radiomics, Preprocessors and classifiers, Colorectal cancer, Computed tomography

Colorectal cancer (CRC) is a frequent malignant tumor worldwide, ranking third in the number of new cancer cases and deaths in the United States in 2023¹. Approximately 15% of patients with CRC exhibit a distinct molecular phenotype known as microsatellite instability (MSI)², which can be categorized into three groups

¹Department of Radiology, Nanjing Drum Tower Hospital Clinical College of Nanjing Medical University, No. 321 Zhongshan Road, Nanjing 210008, Jiangsu Province, China. ²Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., 701 Yunjin Rd, Xuhui District, Shanghai 200232, China. ³These authors contributed equally: Yi Ma and Zhihao Shi. ✉email: 13951791807@163.com; zyzhou@nju.edu.cn

according to the frequency of MSI occurrence: microsatellite stable (MSS), low-frequency MSI (MSI-L), and high-frequency MSI (MSI-H)³. Accurate determination of MSI status is critical in guiding clinical treatment strategies, as reflected in the diagnostic and treatment guidelines of CRC that recommend MSI testing for all patients with CRC.

The MSI status has important implications for the diagnosis, treatment reaction, and prognostic result of CRC. First, MSI is the characteristic molecule of the most common hereditary CRC syndrome, also known as Lynch syndrome. Thus, the MSI status identifies the families with this syndrome and reminds their risk of the disease⁴. Second, patients with MSI are more likely to derive therapeutic benefits from programmed death receptor1 (PD-1) monoclonal antibody treatment other than the traditional fluorouracil chemotherapy⁵⁻⁷. The underlying reason may be related to the more obvious cancer cell mutations and easier immune recognition in patients with MSI^{8,9}. Third, the 5-year survival rate of CRC patients with MSI is significantly longer than that of patients with MSS, particularly in patients with stages II and III CRC¹⁰.

The most common detection methods for MSI include immunohistochemistry (IHC) and polymerase chain reaction (PCR). Both methods are invasive and have high inspection costs^{11,12}. Surgical resection is the ideal method of obtaining histological specimens for testing¹³. However, it is not recommended clinically for lesions with distant metastasis. Considering the impact of tumor heterogeneity, the small portion of sample tissue obtained through biopsy may not accurately reflect the MSI status¹⁴. Repeated biopsies may increase the risk of tumor bleeding, dissemination, etc. Therefore, it is necessary to develop a non-invasive, economical, and effective preoperative prediction method for MSI.

Radiomics can obtain many informative features that cannot be observed by the naked eye from conventional images. Further, it can not only non-invasively and quantitatively evaluate tumor heterogeneity, but also deeply mine the clinicopathological information contained in big data, providing more objective and accurate support for clinical decision-making^{15,16}. Presently, radiomics has been widely used in preoperative diagnosis^{17,18}, treatment response^{19,20}, and prognostic assessment of CRC^{21,22}. Several studies have confirmed that the radiomics features of enhanced CT could help identify preoperative MSI status in CRC patients²³⁻²⁶. However, during the model construction, only two studies mentioned the use of preprocessors, and both used one type^{23,26}. In addition, they all used only one type of classifier. Data processing is crucial in machine learning, different preprocessors and classifiers have different data processing functions, which may affect the performance and generalization ability of the models^{27,28}. Therefore, it's necessary to choose the appropriate preprocessor and classifier for improving the model performance.

Our study retrospectively collected the clinicopathological data of CRC patients. Based on six preprocessors and three classifiers, multiple discriminant models, a clinical screening model, and a nomogram were constructed to predict MSI status. We aimed to explore and optimize the combination of multiple preprocessors and classifiers to improve the performance and generalization ability of prediction models.

Materials and methods

Patients and data

The ethics review committee of Nanjing Drum Tower Hospital approved this retrospective study and waived the informed consent form. All the procedures involving human participants were followed in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The data of patients with CRC confirmed by surgery and pathology in our hospital were collected continuously from January 2020 to October 2022. The inclusion criteria were as following: (1) before surgery the patients received abdominal enhanced computed tomography (CT) examination, (2) pathologically confirmed CRC, and (3) MSI status tests by IHC were available. The exclusion criteria were as following: (1) the interval between CT scan and surgery were more than 2 weeks ($n=15$), (2) insufficient image quality to distinguish tumor contour due to motion or metal artifacts ($n=18$), and (3) any anti-tumor treatment before CT scan ($n=32$). Figure 1 presents the specific inclusion and exclusion criteria.

The collected clinical and pathological indicators included history with or without hypertension, diabetes, sex, age, tumor location, and the clinical of TNM stage. Tumor markers, including CEA, CA125, and CA199, were the results of the last laboratory examination before operation. These results were confirmed by two clinicians.

MSI status assessment

The pathological tissues were stained during IHC using the standard streptavidin–biotin peroxidase process²⁹. Subsequently, the status of MSI was identified by assessing the IHC staining results of four major mismatch repair (MMR) proteins (MLH1, PMS2, MSH2, and MSH6) contained in the tissue. In the four MMR proteins, any lack of expression was considered as MSI, while all positive expressions were considered as MSS³⁰.

CT scan

All patients were scanned using the same 160-slice CT scanner (uCT 780, United Imaging Healthcare, Shanghai, China). Each patient received an informed consent form at the time of appointment for CT scan, covering unified pre-examination preparation work. It was necessary to fast for more than 4 h before the examination and take 250–300 mL of water orally 30 min before scanning. In order to improve the standardization of examinations, an integrated scanning protocol had been developed specifically for the patients, including a unified scanning sequence package and contrast agent. Omnipaque (350 mg I/mL, GE Healthcare) with a dose of 1.5 mL/kg was administered through the anterior elbow vein using a high-pressure syringe at rate of 2.5–3.0 mL/s. Each patient underwent plain scanning, followed by three phases of enhanced scanning. Starting from the injection of contrast agent, the triggering of the arterial, venous, and delay phases scans was delayed for 40 s, 70 s, and 180 s, respectively. The scanning field was from the diaphragm top to the pubic symphysis level. The parameters

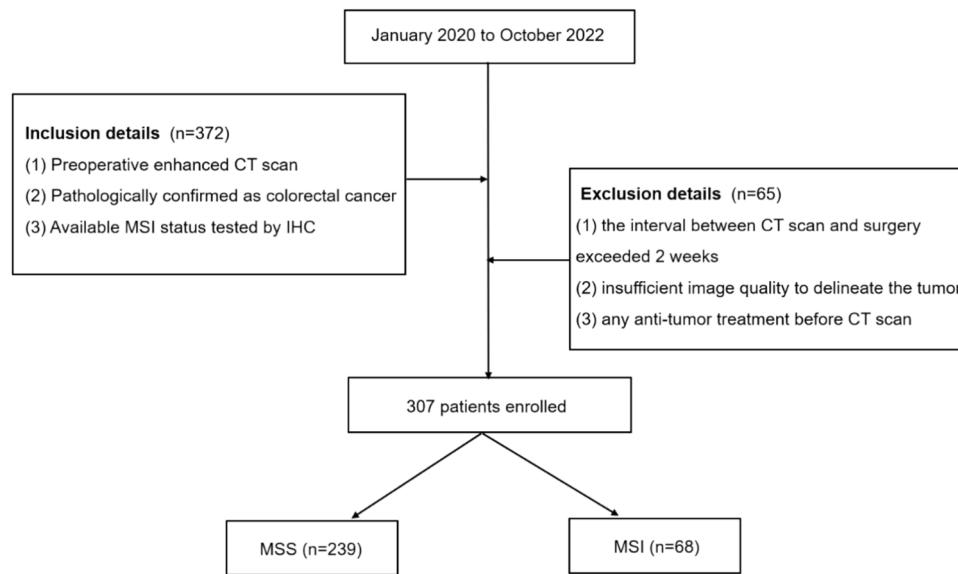


Fig. 1. Patient screening and grouping process. MSS Microsatellite stability, MSI Microsatellite instability, IHC immunohistochemistry.

were as follows: tube current: automatic mAs, tube voltage: 120 kV, pitch: 0.9875: 1, rotation time: 0.5 s, matrix: 512×512 , field of view: 350×350 mm. All of the images were reconstructed with hybrid iterative reconstruction (KARL 3D, United Imaging Healthcare, Shanghai, China) at a 5.0-mm layer thickness and 5.0-mm layer spacing.

Image processing and feature extraction

The venous phase images were selected and sent to the uAI Research Portal software (Shanghai United Imaging Intelligence Co., Ltd.). Its workflow consisted of four parts: image annotation, feature extraction, feature selection, model construction and evaluation (Fig. 2). All tumors were manually drawn by a senior diagnostic radiologist (reader 1 with 11 years of experience), who was blinded to the status of MSI. The cross-section with the largest tumor area was chosen, including necrotic and bleeding areas, while avoiding blood vessels, perienteric

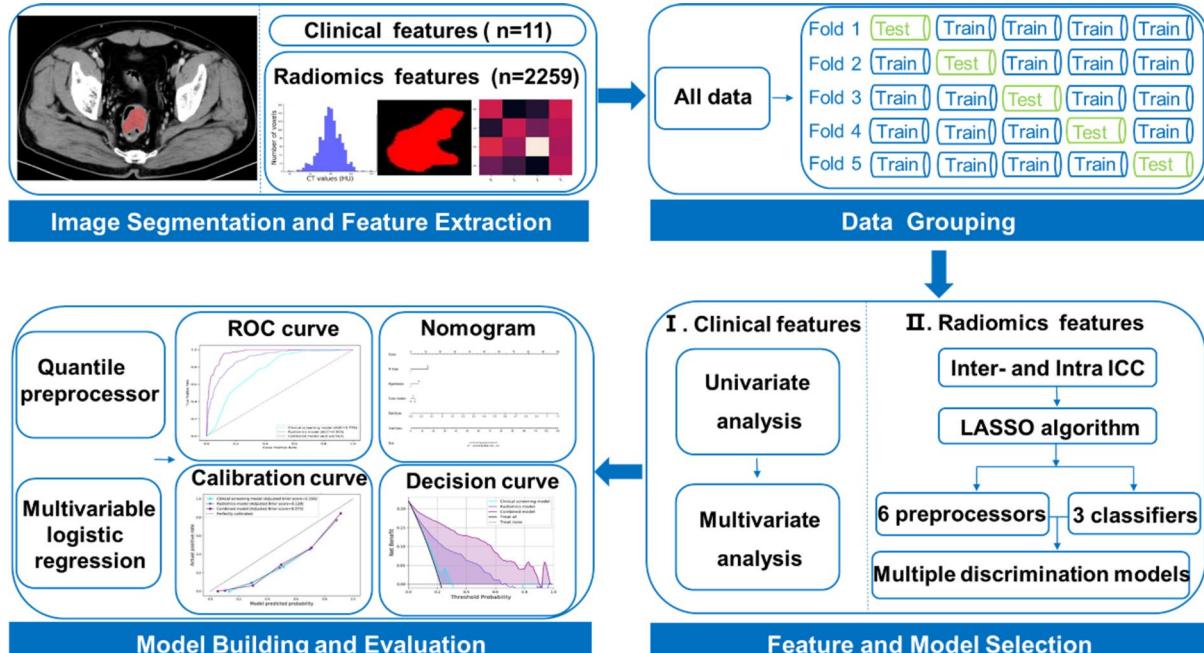


Fig. 2. Workflow of MSI status prediction of colorectal cancer patients including image segmentation and feature extraction, data grouping, feature and model selection, and model building and evaluation.

fat, intestinal contents, and gas. These areas were marked as regions of interest (ROI) (Fig. 3). The largest tumor was chosen to draw the ROI for patients with multiple ones.

Two-dimensional radiomics features were collected from the extensive used radiomics toolbox of PyRadiomics³¹, which contains seven stable feature categories and 14 image filters. Ultimately, 2,259 features were picked up from each ROI. Detailed information on the radiomics features can be obtained in our previous study²⁶.

Feature selection and model construction

After generating the features, machine-learning methods were utilized to select appropriate features and predict the MSI status in CRC patients. To avoid the sample bias of grouping, a stratified fivefold cross-validation strategy was used to randomly but equally divided all the patients into five partitions to make sure that the same percentage of each class (i.e., MSI/MSS) was preserved in each partition. Finally, five different training and test sets were acquired, and the mean value was taken to obtain a more reliable and accurate sample evaluation. To ensure the robustness and generalizability of each model, the feature selection and prediction process was limited to training set, and the parameters obtained from the training cohort were applied to the test set.

Before the feature selection, we first used inter-/intra-correlation coefficients (ICCs) to evaluate inter-/intra-delineator reproducibility. In detail, about two months after the completion of the image delineation, 30 patients³² were randomly selected, and the above steps were finished by reader 1 and another radiodiagnosis physician (reader 2 with 8 years of experience) to segment the images, i.e., manually delineate the ROIs of 30 patients and extract the radiomics features. Features with ICCs less than 0.75 were excluded. Subsequently, the least absolute shrinkage and selection operator algorithm (LASSO) was used to pick the most predictive feature subset within each training set of the fivefold cross-validation. The corresponding coefficients of the selected features were evaluated and utilized to calculate each patient's Rad-score. The Rad-Score of each sample in the test set was computed based on the LASSO coefficients of the corresponding training set and the feature values of the test set sample itself. The following equation was used to calculate the Rad-score:

$$\text{Rad-score} = \sum_{i=1}^n C_i \times X_i + b$$

where n is the number of selected features, C_i is the coefficient of the i th feature from the LASSO regression algorithm, X_i is the i th feature, and b is the intercept of LASSO.

Based on six feature preprocessors (Box-Cox, Yeo-Johnson, Max-Abs, Min-Max, Z-score, and Quantile) and three classifiers [logistic regression, support vector machine (SVM), and random forest], different discriminant models were constructed in the training set using the screened radiomics features. Logistic regression is a well-established and interpretable method, suitable for linear relationship problems³³. SVM is known for its ability to handle complex data patterns and nonlinear relationships or when the decision boundaries are not linearly separable³⁴. Random forest, an ensemble learning method, offers robustness and good performance through the combination of multiple decision trees³⁵. These classifiers have been widely used and demonstrated effectiveness in studies^{36–38}, making them suitable choices for our analysis.

In the test stage, the trained models were applied to the test dataset to predict the probability of being MSI or MSS status. The model with the highest average value of the area under the curve (AUC) in the test set was chosen as the radiomics model. To predict the MSI status, multivariate regression analysis was performed on clinical characteristics with P values less than 0.1 in the difference analysis to screen out the clinical independent factors. The same feature preprocessing algorithm and classifier of the radiomics model were used to develop the clinical screening model and combined model. The clinical screening model was composed of clinical independent factors, whereas the combined model, including the clinical independent factors and the Rad-score derived from the LASSO feature selection process. To provide clinicians a convenient and user-friendly approach for rapidly and accurately estimating the risk of MSI status in individual patients, a nomogram model was developed. It should be noted that all available data was employed for training and estimating the parameters of the nomogram,



Fig. 3. The tumor with the largest area in cross-section were segmented on venous phase, avoiding the intestinal contents and gas.

which allows for a more comprehensive understanding of the overall patterns and relationships. Specifically, the clinical characteristics and Rad-score values were directly obtained by concatenating the test sets from the fivefold cross-validation used in the construction of combined model. Additionally, three features were randomly selected from the features screened by LASSO to perform six data transformations to compare the feature processing results of different preprocessors. In model construction, the hyperparameters were defined using the training set with a grid search to optimize predictive accuracy, detailed information can be found in the Supplementary material.

Statistical analysis

We separately used the Mann–Whitney U and the χ^2 test to compare the continuous and the categorical variables. The statistical analyses were bilateral, and statistical difference was set to $P < 0.05$. To evaluate and verify the predictive effectiveness of the models, the receiver operating characteristic (ROC) curves of the clinical, radiomics, and combined models were analyzed, respectively. We used the DeLong test to statistically compare the AUC values obtained from the different prediction models. The average performance of each model was evaluated across the fivefold cross-validation. The clinical applicability and correction effects of the models were compared using decision curve analysis (DCA) and calibration curves. The Brier score (BS) was used to calculate the quantitative analysis of each model performance: BS = 0 indicates that the model performs excellently and the predicted and actual values were identical; BS > 0.25 implies the failure of the model prediction. To address the impact of class imbalance on our calibration curves analysis, the BS value was adjusted based on the class distribution. All statistical tests were executed using IBM SPSS Statistics for Windows, version 26 (IBM Corp., Armonk, N.Y., USA) and R software (version 3.5.2; <http://www.Rproject.org>). All feature preprocessing and model construction were carried out using the scikit-learn package in Python 3.9.12.

Results

Patient profiles

There were 307 CRC patients with 182 males (59.3%) and 125 females (40.7%) were enrolled. Their average age was 62.7 ± 12.0 years (27–93 years), including 68 (22.1%) patients with MSI, and 239 (77.9%) patients with MSS. In Table 1, the clinical analysis of the two groups showed statistical differences in hypertension ($P = 0.009$), the clinical of N stage ($P < 0.001$), and tumor location ($P < 0.001$). The multivariate regression analysis identified

Variable	MSI (n = 68)	MSS (n = 239)	P value
Gender			0.077
Male	34	148	
Female	34	91	
Age (years)	64 (53, 73)	63 (55, 71)	0.741
Hypertension			0.009
Presence	16 (23.53%)	98 (41.00%)	
Absence	52 (76.47%)	141 (59.00%)	
Diabetes			0.196
Presence	11 (16.18%)	25 (10.46%)	
Absence	57 (83.82%)	214 (89.54%)	
Tumor location			< 0.001
Right colon	43 (63.23%)	75 (31.38%)	
Left colon	13 (19.12%)	45 (18.83%)	
Rectum	12 (17.65%)	119 (49.79%)	
T stage			0.707
T1 ~ 2	9 (13.24%)	36 (15.06%)	
T3 ~ 4	59 (86.76%)	203 (84.94%)	
N stage			< 0.001
N0	55 (80.88%)	113 (47.28%)	
N1 ~ 2	13 (19.12%)	126 (52.72%)	
M stage			0.344
M0	68 (100.00%)	233 (97.49%)	
M1	0 (0.00%)	6 (2.51%)	
CEA (ng/ml)	2.22 (0.94, 4.13)	2.65 (1.43, 6.46)	0.074
CA125 (U/ml)	8.95 (6.20, 16.48)	8.00 (5.10, 11.30)	0.061
CA199 (U/ml)	15.65 (7.44, 25.98)	12.26 (7.15, 30.49)	0.483

Table 1. Characteristics of patients [median (Q1, Q3) or no. (%)]. P value was derived from the Mann–Whitney U test or the χ^2 test. P value less than .05 were considered as statistically significant. CEA: carcinoembryonic antigen level; CA: carbohydrate antigen.

hypertension [OR 0.378 (95% confidence intervals (CI), 0.191–0.748), $P=0.005$], N stage [OR 0.195 (95% CI 0.096–0.395), $P<0.001$], and tumor location [OR 0.347 (95% CI 0.139–0.866), $P=0.023$] as independent factors of MSI status.

Model building and processor analysis

We evaluated the average performance of each model across the fivefold cross-validation. In Table 2, we can see that under different types of preprocessors, the models established by the logistic regression classifier all had the higher average AUC value. And the logistic regression model based on the quantile preprocessor had the highest average AUC value of 0.852 [95% confidence interval (CI) 0.750–0.958] in all the discriminant models (Table S2). It was selected as the radiomics model which included 23 radiomic features (Figure S1). The logistic regression model based on the quantile preprocessor was also used to build the clinical screening model as well as the combined model. In the test cohort, the clinical screening model produced moderate performance with an average AUC value of 0.762 (95% CI 0.635–0.890), and the combined model yielded the excellent performance with an average AUC value of 0.958 (95% CI 0.920–0.998) (Table 3). In order to assess their predictive performance, the average ROC curves of the training and test sets were presented in Fig. 4. To provide a more comprehensive assessment of the model's performance across different training and test sets on the fivefold cross-validation, the predictive performance and ROC curves for each fold of the combined model was presented in Table S4 and Figure S2, respectively. The results revealed that each fold achieved a desirable diagnostic performance, with AUC

Feature selection methods	Box-Cox		Max-Abs		Min-Max		Quantile		Yeo-Johnson		Z-score	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Logistic regression	0.914	0.846	0.907	0.839	0.893	0.833	0.903	0.852	0.911	0.845	0.911	0.845
SVM	0.913	0.837	0.981	0.823	0.909	0.824	0.906	0.827	0.906	0.829	0.906	0.829
Random forest	0.965	0.775	0.942	0.801	0.941	0.803	0.973	0.735	0.956	0.786	0.953	0.780

Table 2. Analysis of average AUC values for 18 discriminant models with preprocessing. AUC, area under the curve. SVM, support vector machine.

Cohorts	AUC (95% CI)			<i>P</i> value (0 vs. 1)	<i>P</i> value (0 vs. 2)	<i>P</i> value (1 vs. 2)
	Clinical screening model (0)	Radiomics model (1)	Combined model (2)			
Training set	0.770 (0.708–0.832)	0.903 (0.863–0.944)	0.963 (0.946–0.985)	<0.001	<0.001	<0.001
Test set	0.762 (0.635–0.890)	0.852 (0.750–0.958)	0.958 (0.920–0.998)	0.030	<0.001	<0.001

Table 3. Pairwise comparisons of average AUCs of the clinical screening model, radiomics model, and combined model. *P* value derived from Delong test.

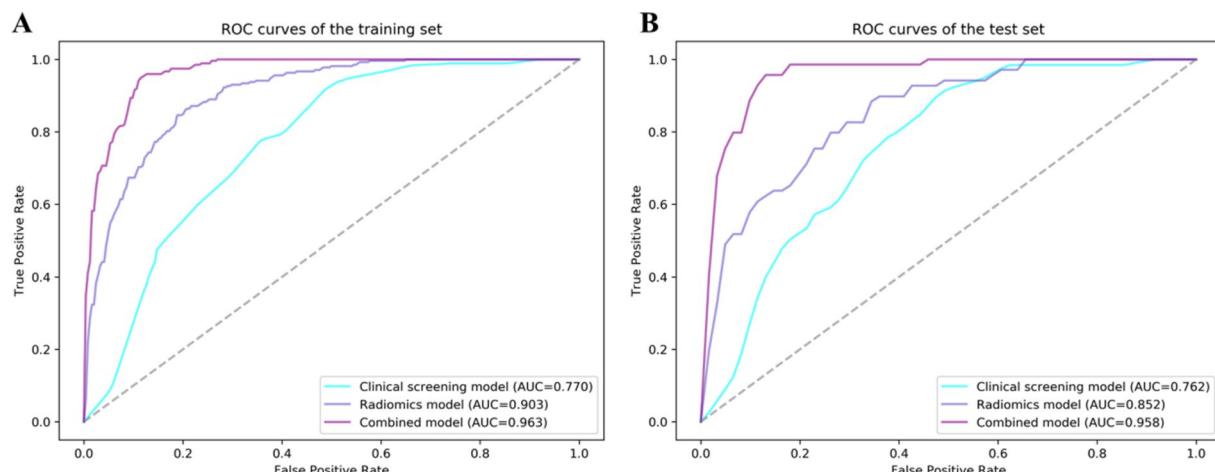


Fig. 4. The average receiver operating characteristic (ROC) curves of the clinical screening model, radiomics model and combined model in the training set (A) and test set (B). The combined model performed better than the other two models with the average area under the curve (AUC) of 0.963 and 0.958 in the training and test set, respectively.

values ranging from 0.959 to 0.978 and accuracy ranging from 0.886 to 0.910 in the training set. Similarly, in the test set, the AUC values ranged from 0.912 to 0.987 and accuracy ranged from 0.855 to 0.934. After the processing of the six feature preprocessors, the result data of the quantile transformer tended to be similar to a normal distribution, and the mean value of the MSS features was higher than that of MSI (Fig. 5). Additionally, ablation experiments were conducted on the selected radiomics features without preprocessing (Table S2). The results showed that the use of preprocessing improved average AUC values of the discriminative models by at least 2%.

Clinical application

To promote clinical practice, a quantitative nomogram was developed for non-invasive prediction of MSI status (Fig. 6). We used the consistency indexes (C-indexes) to estimate its performance. The C-index of the nomogram reached 0.970, indicating its excellent effectiveness. The average predictive ability of the combined model (accuracy: 0.899; sensitivity: 0.929; specificity: 0.891) was superior to that of the radiomics and clinical screening model in the test set (Table S3). The Delong test further revealed statistical differences between any two models

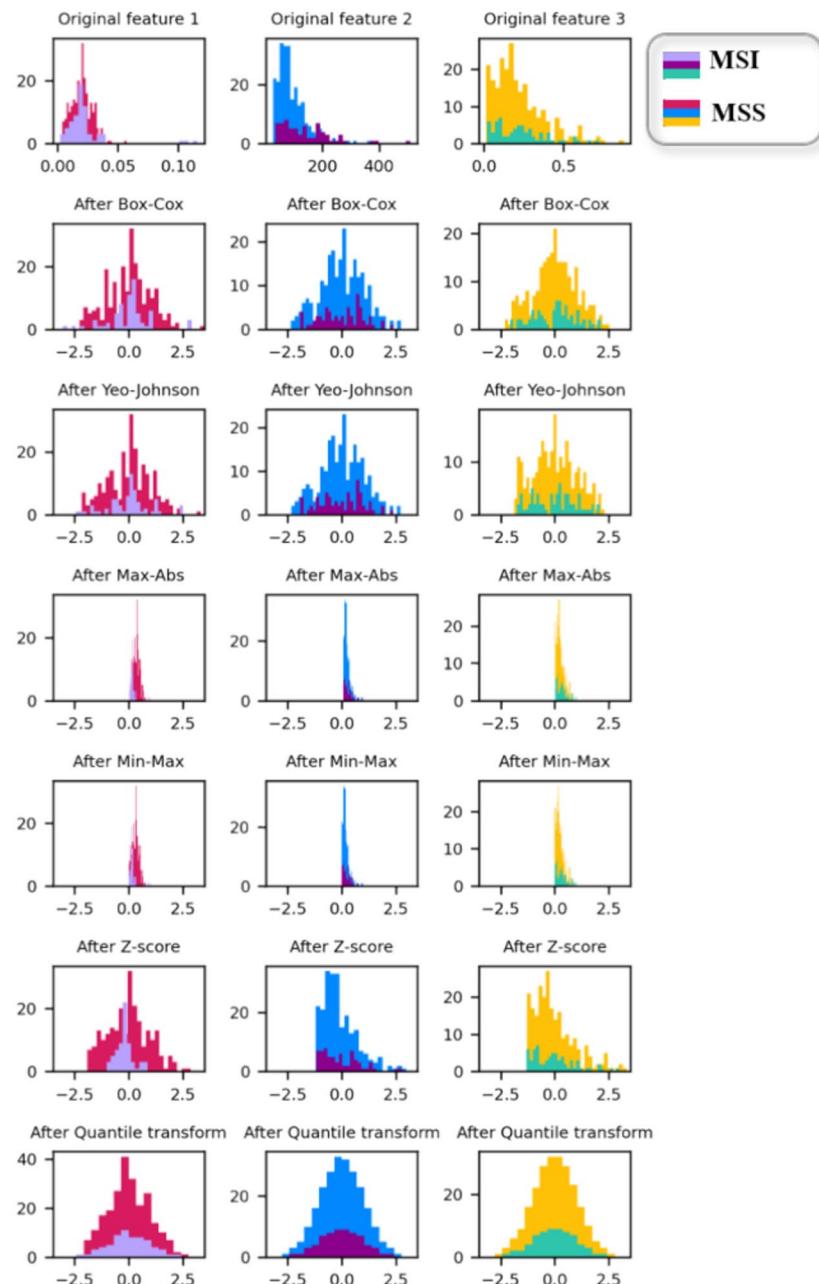


Fig. 5. Randomly selected 3 features for six data transformations to compare the feature processing results of different preprocessors. The results showed that the data processed by the Quantile transformer were closer to the normal distribution, which can make better distinguish between MSI and MSS.

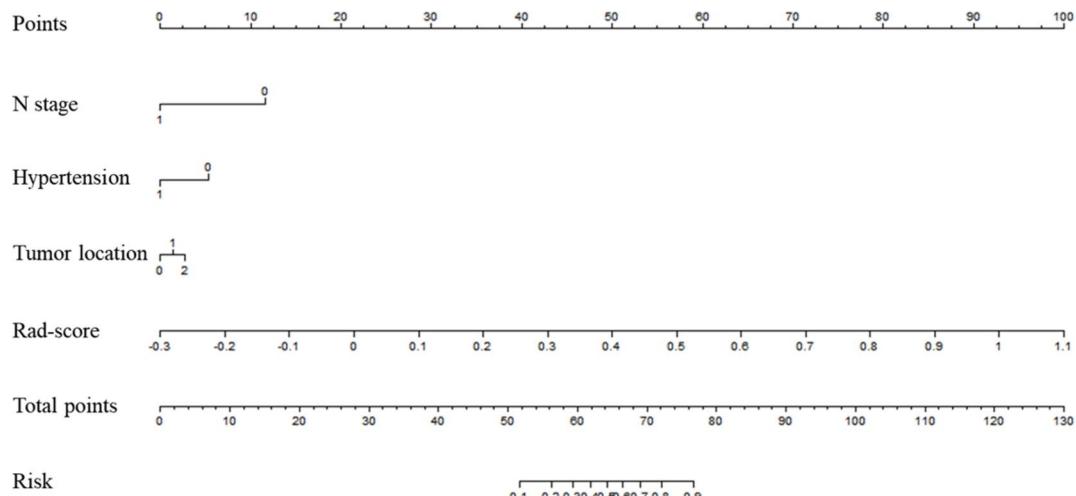


Fig. 6. An individualized nomogram for preoperative prediction of MSI status in patients with colorectal cancer. In terms of the N stage, the value of 0 indicates N0 while 1 represents either N1 or N2. For hypertension, 0 indicates absence of hypertension, whereas 1 indicates that the patient has hypertension. Regarding the location, 0 corresponds to the left colon, 1 indicates the right colon, and 2 refers to the rectum. As for Rad-score, the value was calculated for each patient through a linear combination of the selected features weighted by their corresponding LASSO coefficients. Note that, all the test sets data from the fivefold cross-validation of combined model construction were concatenated and used for the nomogram model development. When using the nomogram, first locate each variable on the axis, and then draw a vertical line towards the points axis to obtain the corresponding score. Finally, by summing all the scores and positioning them on the bottom line, the predicted incidence of MSI states could be obtained.

($P < 0.05$) (Table 3). It indicated that the combined model can better predict MSI status than the radiomics model and the clinical screening model in the calibration curve. The adjusted BS values of the clinical screening model, radiomics model, and combined model were 0.196, 0.128, and 0.073 in the training set and 0.199, 0.164, and 0.079 in the test set, respectively (Fig. 7A,B). DCA revealed that compared to the other two models, the combined model generally had the best net benefit value over the entire range (Fig. 7C,D).

Discussion

In our study, six preprocessors and three classifiers were used to build models to predict the status of MSI in CRC. It revealed that the logistic regression model based on the quantile preprocessor exhibited good predictive performance. The same combination was also used to build the clinical and combined models. The results showed that the clinical screening model demonstrated moderate predictive performance, with an average AUC value of 0.762 (95% CI 0.635–0.890) in the test cohort. Furthermore, the combined model demonstrated excellent predictive ability with an average AUC value of 0.958 (95% CI 0.920–0.998). This further confirms the performance and repeatability of the chosen combination.

In this study, 11 clinical factors were included. MSI incidence was 22.15% (68/307). It mainly happened in the right colon, and the incidence rate was 63.23% (43/68), consistent with previous studies^{39,40}. Aside from tumor location, we discovered that the clinically independent factors to predict MSI status also included N-stage and hypertension. Lymph node metastasis is an important prognostic factor for CRC, the higher the N-stage, the shorter the patient's survival period maybe^{41,42}. In addition, the patients with metabolic syndrome such as hypertension or diabetes may have a higher risk with disease recurrence and death⁴³. The previous reports^{10,44} have confirmed that CRC patients with MSI usually have a favorable outcome. It may be related to the lower incidence rate of lymph node metastasis and hypertension incidence in our study. However, a recent research⁴⁵ based on 100 patients found that there was no significant relationship between hypertension and MSI status. These results need to be confirmed by future studies with larger sample sizes.

Radiomics analysis can extract high-throughput features hidden in images to reflect tumor heterogeneity⁴⁶. It has been widely used in the field of prognosis, treatment evaluation, and survival prediction of many clinical diseases^{47,48}. Pathologically, the histological heterogeneity of MSI CRC is more obvious than that of MSS. The former has a higher proportion of lymphocyte infiltration and mucus components⁴⁹. These histological differences have laid the foundation for radiomics analysis.

Previous studies have predominantly used a single preprocessor and classifier to establish predictive models. For instance, Cao et al.²³ and our previous work²⁶ both used the preprocessor of Z-score normalization and the classifier of logistic regression to build the models. The combined models showed excellent predictive ability with an AUC of 0.964 (95% CI: 0.919–1.000) and 0.928 (95% CI: 0.860–0.991) in the validation cohort, respectively. Ying et al.²⁴ and Pei et al.²⁵ proposed the combined model with an AUC of 0.900 (95% CI 0.830–0.960) and 0.770 (95% CI 0.680–0.850) in the validation sets. They both used the logistic regression classifier, but the preprocessor did not explicitly state. As we know, data processing is crucial in machine learning. Through appropriate

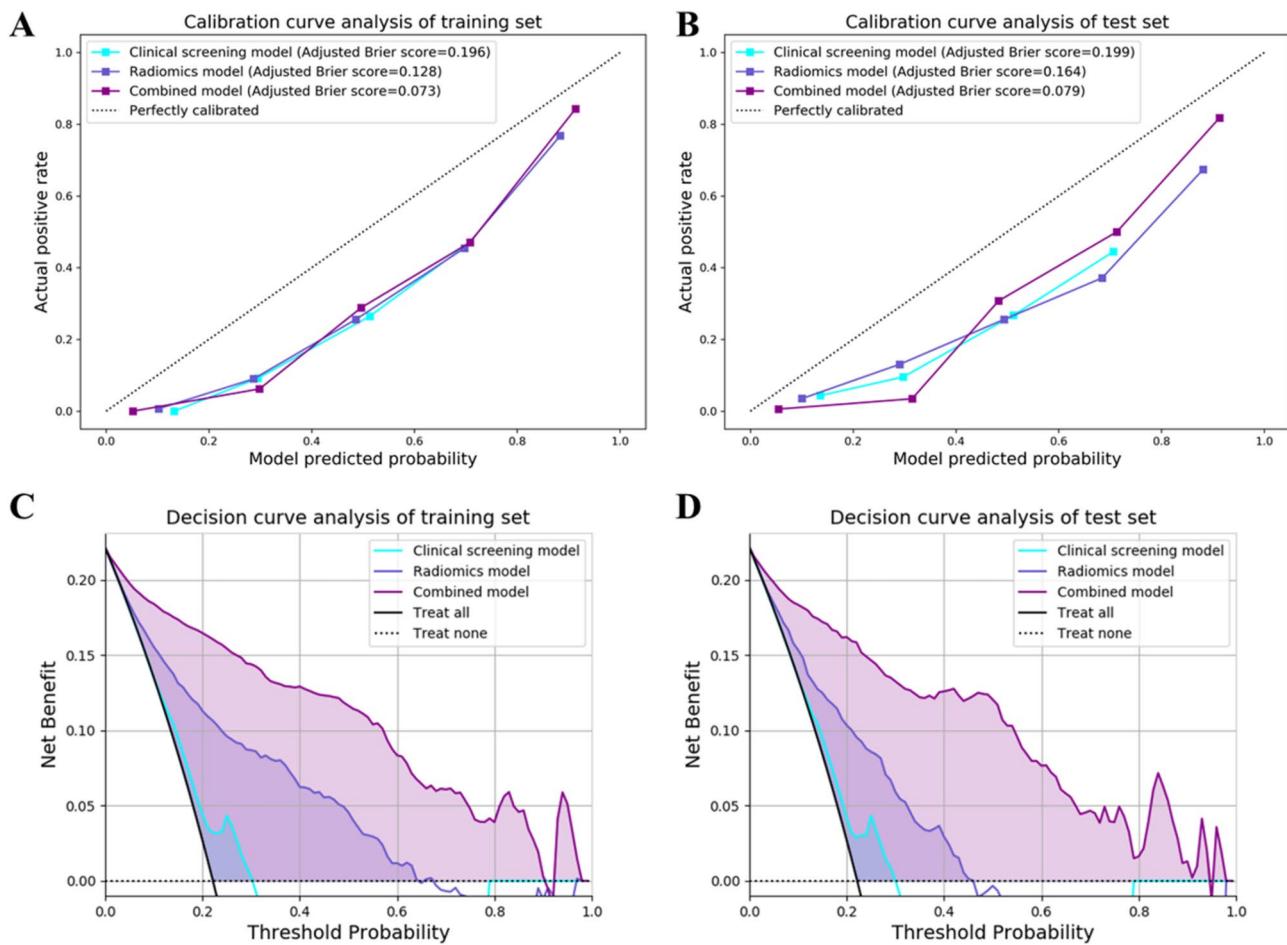


Fig. 7. Calibration curves of the clinical screening model, radiomics model and combined model in the training set (A) and test set (B). The dotted line represented a perfect prediction, while the solid lines with three different colors indicated the predictive performance of the three models. The closer to the dotted line, the better prediction of the model. The Brier score values were adjusted based on the imbalanced class distribution. Calibration curves showed the prediction performance for MSI status of the combined model was generally better than the other two in the test set. Decision curve analysis (DCA) of the three model (C–D). The x-axis was the threshold probability and the y-axis was the net benefit. Within any threshold probability range, a higher curve was the best prediction for maximizing the net benefit. The DCA showed that the combined model had the highest net benefit almost across the entire range.

preprocessing, the raw data can be transformed into data features for model use. Subsequently, the classifier can classify the filtered features into reasonable categories to improve the performance and generalization ability of the model^{27,28}. Therefore, our study simultaneously selected 6 preprocessors and 3 classifiers to constructed the models, and explored the impact of different combinations on model performance.

Consistent with the recent studies^{23–26}, we selected the logistic regression classifier to establish the predictive models. The logistic regression model is suitable for solving binary classification problems in machine learning, which can quickly learn and predict data, and the obtained results are easy to explain⁵⁰. These characteristics make it advantageous in predicting MSI status. However, there were several scenarios where SVM or random forest could potentially outperform logistic regression. For example, in image recognition tasks or when dealing with datasets with high-dimensional feature spaces and intricate patterns, SVM often shows superior performance³⁴. Random forest, on the other hand, tends to perform better when there are numerous features and potential interactions among them. It's also more robust to noise and outliers in the data³⁵. In our framework, including these three classifiers allows us to comprehensively evaluate and compare their performance, and thus determine which classifier is more suitable for our given problem.

Additionally, as shown in Table S2, the preprocessors play a relatively minor role in influencing the classification performance, indicating that they mainly focus on operations such as data cleaning, normalization, and similar tasks that do not substantially alter the fundamental nature and discriminatory power of the features. On the contrary, the type of features and the choice of classifier have a more significant impact on the classification performance. Different feature types capture distinct aspects of the data, and each classifier has its inherent strengths and weaknesses in handling and learning from these features. Regarding our task, the combination of the logistic regression classifier and the quantile transformer proved to be effective in predicting

for the MSI status of CRC patients, highlighting the importance of selecting the right combination of algorithm to optimize classification results.

To facilitate clinical applications, a nomogram was developed to optimize treatment strategies. Our study screened common clinical indicators such as location, N-stage, and hypertension as predictive factors. By constructing a nomogram, it was expected to save the cost of preoperative individualized and precise prediction of MSI status. What's more, the effectiveness and repeatability of the selected processor and classifier have been preliminarily validated in the construction of clinical screening model and combined model. Standardized processing can further reduce the variability of the input data, promote the homogenization of different researches and ensure comparability of results.

However, our research still had several limitations. Firstly, it was a single-center study with a limited sample size, it will be necessary to further validate our results through external and multicenter studies. Secondly, we chosen the largest layer of tumor to draw the ROI, similar to previous studies^{32,51}. But it may result in a certain degree of selection bias. Thirdly, all of the CT images were obtained from a same scanner, it may affect the generalizability of our findings, although it reduced changes in image acquisition.

Conclusion

Simultaneously using multiple preprocessors and classifiers to construct predictive models, our results showed that the logistic regression model based on the quantile preprocessor exhibited excellent predictive performance and repeatability. It may further reduce the variability of input data and improve the model performance for predicting MSI status in CRC.

Data availability

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

Received: 20 February 2024; Accepted: 28 August 2024

Published online: 01 September 2024

References

1. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **73**(1), 17–48 (2023).
2. Bonneville, R. *et al.* Landscape of microsatellite instability across 39 cancer types. *JCO Precis Oncol.* **1**, 1 (2017).
3. De' Angelis, G. L. *et al.* Microsatellite instability in colorectal cancer. *Acta Biomed.* **89**(9-S), 97–101 (2018).
4. Vasen, H. F. Clinical description of the Lynch syndrome [hereditary nonpolyposis colorectal cancer (HNPCC)]. *Fam. Cancer.* **4**(3), 219–225 (2005).
5. Ribic, C. M. *et al.* Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *N. Engl. J. Med.* **349**(3), 247–257 (2003).
6. Chalabi, M. *et al.* Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers. *Nat. Med.* **26**(4), 566–576 (2020).
7. Le, D. T. *et al.* PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**(26), 2509–2520 (2015).
8. Fischer, F., Baerenfaller, K. & Jiricny, J. 5-Fluorouracil is efficiently removed from DNA by the base excision and mismatch repair systems. *Gastroenterology.* **133**(6), 1858–1868 (2007).
9. Hegde, M., Ferber, M., Mao, R., Samowitz, W. & Ganguly, A. Working Group of the American College of Medical Genetics and Genomics (ACMG) Laboratory Quality Assurance Committee. ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). *Genet. Med.* **16**(1), 101–116 (2014).
10. Merok, M. A. *et al.* Microsatellite instability has a positive prognostic impact on stage II colorectal cancer after complete resection: results from a large, consecutive Norwegian series. *Ann. Oncol.* **24**(5), 1274–1282 (2013).
11. Sepulveda, A. R. *et al.* Molecular Biomarkers for the Evaluation of Colorectal Cancer: Guideline From the American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and the American Society of Clinical Oncology. *J. Clin. Oncol.* **35**(13), 1453–1486 (2017).
12. Umar, A. *et al.* Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J. Natl. Cancer Inst.* **96**(4), 261–268 (2004).
13. Harada, S. & Morlote, D. Molecular pathology of colorectal cancer. *Adv. Anat. Pathol.* **27**(1), 20–26 (2020).
14. Sacher, A. G. *et al.* Association between younger age and targetable genomic alterations and prognosis in non-small-cell lung cancer. *JAMA Oncol.* **2**(3), 313–320 (2016).
15. Aerts, H. J. The potential of radiomic-based phenotyping in precision medicine: A review. *JAMA Oncol.* **2**(12), 1636–1642 (2016).
16. Gardin, I. *et al.* Radiomics: Principles and radiotherapy applications. *Crit. Rev. Oncol. Hematol.* **138**, 44–50 (2019).
17. Yin, J. D., Song, L. R., Lu, H. C. & Zheng, X. Prediction of different stages of rectal cancer: Texture analysis based on diffusion-weighted images and apparent diffusion coefficient maps. *World J. Gastroenterol.* **26**(17), 2082–2096 (2020).
18. Xue, T. *et al.* Preoperative prediction of KRAS mutation status in colorectal cancer using a CT-based radiomics nomogram. *Br. J. Radiol.* **95**(1134), 20211014 (2022).
19. Cheng, Y. *et al.* Multiparametric MRI-based radiomics approaches on predicting response to neoadjuvant chemoradiotherapy (nCRT) in patients with rectal cancer. *Abdom. Radiol. (NY)* **46**(11), 5072–5085 (2021).
20. Shin, J. *et al.* MRI radiomics model predicts pathologic complete response of rectal cancer following chemoradiotherapy. *Radiology.* **303**(2), 351–358 (2022).
21. Lovinfosse, P. *et al.* FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer. *Eur. J. Nucl. Med. Mol. Imaging.* **45**(3), 365–375 (2018).
22. Xue, T. *et al.* A CT-based radiomics nomogram in predicting the postoperative prognosis of colorectal cancer: A two-center study. *Acad. Radiol.* **29**(11), 1647–1660 (2022).
23. Cao, Y. *et al.* Predicting microsatellite instability status in colorectal cancer based on triphasic enhanced computed tomography radiomics signatures: A multicenter study. *Front. Oncol.* **11**, 687771 (2021).
24. Ying, M. *et al.* Development and validation of a radiomics-based nomogram for the preoperative prediction of microsatellite instability in colorectal cancer. *BMC Cancer.* **22**(1), 524 (2022).
25. Pei, Q. *et al.* Pre-treatment CT-based radiomics nomogram for predicting microsatellite instability status in colorectal cancer. *Eur. Radiol.* **32**(1), 714–724 (2022).

26. Ma, Y. *et al.* Radiomics features based on internal and marginal areas of the tumor for the preoperative prediction of microsatellite instability status in colorectal cancer. *Front. Oncol.* **12**, 1020349 (2022).
27. Dewi, D. E. O. *et al.* The impact of pre-processing and disease characteristics on reproducibility of T2-weighted MRI radiomics features. *MAGMA* **36**(6), 945–956 (2023).
28. Zhang, Z. & Sejdic, E. Radiological images and machine learning: Trends, perspectives, and prospects. *Comput. Biol. Med.* **108**, 354–370 (2019).
29. Lindor, N. M. *et al.* Immunohistochemistry versus microsatellite instability testing in phenotyping colorectal tumors. *J. Clin. Oncol.* **20**(4), 1043–1048 (2002).
30. Gelsomino, F., Barbolini, M., Spallanzani, A., Pugliese, G. & Cascinu, S. The evolving role of microsatellite instability in colorectal cancer: A review. *Cancer Treat. Rev.* **51**, 19–26 (2016).
31. Van Griethuysen, J. J. M. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**(21), e104–e107 (2017).
32. Liu, D. *et al.* A bounding Box-Based radiomics model for detecting occult peritoneal metastasis in advanced gastric cancer: A multicenter study. *Front. Oncol.* **11**, 77760 (2021).
33. Lever, J., Krzywinski, M. & Altman, N. Logistic regression. *Nat. Methods* **13**, 541–542 (2016).
34. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **24**(12), 1565–1567 (2006).
35. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
36. Huang, S. *et al.* Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genom. Proteom.* **15**(1), 41–51 (2018).
37. Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**(1), 3133–3181 (2014).
38. Yang, F. *et al.* Machine learning for histologic subtype classification of non-small cell lung cancer: A retrospective multicenter radiomics study. *Front Oncol.* **10**, 608598 (2021).
39. Shin, U. S. *et al.* Is microsatellite instability really a good prognostic factor of colorectal cancer? *Ann. Coloproctol.* **30**(1), 28–34 (2014).
40. Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**(6), 2073–2087.e3 (2010).
41. Kwon, T. S. *et al.* Novel methods of lymph node evaluation for predicting the prognosis of colorectal cancer patients with inadequate lymph node harvest. *Cancer Res. Treat.* **48**(1), 216–224 (2016).
42. Li, M. *et al.* A clinical-radiomics nomogram for the preoperative prediction of lymph node metastasis in colorectal cancer. *J. Transl. Med.* **18**(1), 46 (2020).
43. Yang, Y. *et al.* Effect of metabolic syndrome and its components on recurrence and survival in colon cancer patients. *Cancer* **119**(8), 1512–1520 (2013).
44. Taieb, J. *et al.* Prognosis of microsatellite instability and/or mismatch repair deficiency stage III colon cancer patients after disease recurrence following adjuvant treatment: Results of an ACCENT pooled analysis of seven studies. *Ann. Oncol.* **30**(9), 1466–1471 (2019).
45. Atef, N. *et al.* Microsatellite instability and life style factors in sporadic colorectal cancer. *Asian Pac. J. Cancer Prev.* **21**(5), 1471–1480 (2020).
46. Rogers, W. *et al.* Radiomics: From qualitative to quantitative imaging. *Br. J. Radiol.* **93**(1108), 20190948 (2020).
47. Wang, T. *et al.* Radiomics for survival risk stratification of clinical and pathologic stage IA pure-solid non-small cell lung cancer. *Radiology* **302**(2), 425–434 (2022).
48. Nketiah, G. *et al.* T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results. *Eur. Radiol.* **27**(7), 3050–3059 (2017).
49. Fuchs, T. L. *et al.* Assessment of tumor-infiltrating lymphocytes using international TILs working group (ITWG) system is a strong predictor of overall survival in colorectal carcinoma: A study of 1034 Patients. *Am. J. Surg. Pathol.* **44**(4), 536–544 (2020).
50. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
51. Beig, N. *et al.* Perinodular and intranodular radiomic features on lung CT images distinguish adenocarcinomas from granulomas. *Radiology* **290**(3), 783–792 (2019).

Author contributions

Study concept and design: Z.Y.Z., G.C.Q. Statistical analyses: Y.W., F.S. data curation, Z.H.S. writing—original draft preparation, Y.M. writing—review and editing Z.Y.Z. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-71420-4>.

Correspondence and requests for materials should be addressed to G.Q. or Z.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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

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

© The Author(s) 2024