Heliyon 11 (2025) e41852

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Preoperative prediction of multiple biological characteristics in colorectal cancer using MRI and machine learning

Qiao-yi Huang [a,1], Hui-da Zheng [b,1], Bin Xiong [b], Qi-ming Huang [c], Kai Ye [b], Shu Lin [d,*], Jian-hua Xu [b,**]

[a] Department of Gynaecology and Obstetrics, The Second Affiliated Hospital, Fujian Medical University, Quanzhou, Fujian Province, China
[b] Department of Gastrointestinal Surgery, The Second Affiliated Hospital of Fujian Medical University, Quanzhou, Fujian Province, China
[c] Department of Radiology, The Second Affiliated Hospital of Fujian Medical University, Quanzhou, Fujian Province, China
[d] Centre of Neurological and Metabolic Research, The Second Affiliated Hospital of Fujian Medical University, Quanzhou, Fujian Province, China

ARTICLE INFO

Keywords:
Radiomics
Enhanced magnetic resonance imaging
Colorectal cancer
Biological characteristics
Preoperative evaluation

ABSTRACT

Colorectal cancer (CRC) is the second most prevalent cause of oncological mortality, and its diagnostic and therapeutic decision-making processes is complex. Alteration in molecular characteristic expression is closely related to tumor invasiveness and can serve a novel biomarker for predicting cancer prognosis. In this study, we aimed to construct radiomic models through machine learning to predict the progression of CRC. We collected the clinical, pathological, and magnetic resonance imaging (MRI) data of 136 CRC patients who underwent direct surgical resection. Immunohistochemistry analysis was performed to detect the expression levels of p53, synaptophysin (Syn), human epidermal growth factor receptor 2 (HER2), perineural invasion (PNI), and vascular invasion (VI) expression levels in CRC tissues. After the manual lesion segmentation, 1781 radiomics features were extracted from the transverse T2-weighted image of MRI (T2W-MRI). We employed Spearman's rank correlation coefficient, greedy recursive deletion strategy, minimum redundancy, maximum relevance, least absolute shrinkage, and selection operator regression were utilized to screen for radiological features. Radiomics and clinical models were constructed using the K-nearest neighbor (KNN). The diagnostic efficiencies of the prediction models were evaluated using receiver operating characteristic curves and quantified employing the area under the curve (AUC). Our research results indicate that compared with the single radioactive model, the clinical radiomics model in the validation cohort showed better diagnostic performance, as indicated by the AUC values (p53 = 0.758, Syn = 0.739, HER2 = 0.786, PNI = 0.835, VI = 0.797). Furthermore, the calibration curve and decision curve analyses showed the clinical benefits. In summary, we developed and validated a clinical radiomics model to preoperative prediction of the biological characteristic expression levels of CRC. The findings of this research may offer a promising noninvasive method for evaluating CRC risk stratification and may lay the groundwork for treatment of this disease.

* Corresponding author. No.34 North Zhongshan Road, Quanzhou, Fujian Province, 362000, China.
** Corresponding author. No.34 North Zhongshan Road, Quanzhou, Fujian Province, 362000, China.
E-mail addresses: shulin1956@126.com (S. Lin), xjh630913@126.com (J.-h. Xu).
[1] Co-first author: Qiao-yi Huang and hui-da Zheng contributed equally to this article.

https://doi.org/10.1016/j.heliyon.2025.e41852
Received 5 March 2024; Received in revised form 6 January 2025; Accepted 8 January 2025
Available online 9 January 2025

1. Introduction

Colorectal cancer (CRC) ranks third in incidence rate among malignancies (following breast and lung cancers) and is the second highest cause of oncological mortality worldwide [1]. The primary treatment methods to manage CRC include surgical resection, adjuvant chemotherapy, radiation therapy, targeted therapy, and immunotherapy [2,3]. However, assessing the overall therapeutic response and the risk of recurrence in patients remain challenging. Although tissue biopsy is the gold standard for clinical decision-making, its application in tumour monitoring is limited because of repeated injuries, high surgical costs, and poor patient compliance. Additionally, a single biopsy cannot accurately reflect patient heterogeneity. Therefore, noninvasive diagnoses and reliable prognostic factor identification of CRC are essential for patient management and monitoring of CRC recurrence.

Tumour molecular classification is crucial in CRC prognosis prediction. P53 is a tumour suppressor that induces cell-cycle arrest, DNA repair, and apoptosis under oncogenic or other cellular stress [4,5], and nearly half of CRC patients have p53 gene mutations [6,7]. P53 mutation is associated with tumor progression, liver metastases, and chemoresistance of colorectal tumors [8,9]. Adenocarcinoma with neuroendocrine differentiation (NED) is a special CRC type with a shorter survival period [10,11]. Synaptophysin (Syn), an integral membrane glycoprotein of secretory vesicles, is the most commonly used neuroendocrine cell marker in CRC research and can better describe the impact of NED on CRC prognosis better [12]. Human epidermal growth factor receptor 2 (HER2), a tyrosine kinase family member, is up - regulated in multiple neoplasms, including CRC [13]. Furthermore, HER2 overexpression is significantly correlated with lymphatic metastases, distant metastases, and perineural invasion (PNI) [14]. PNI is a tumour invasion of the nerve and nerve sheaths that negatively affects the recurrence and long-term survival of CRC patients [15]. Vascular invasion (VI) is the vascular wall destruction or invasion by tumour cells, which is a crucial step in the metastatic process and an independent adverse prognostic factor for survival among CRC patients [16,17]. Therefore, the preoperative prediction of these biological characteristics can better assess tumour invasiveness; hence, it requires more optimised detection techniques are required.

These molecular classifications can be identified through colonoscopic biopsy or postoperative detection of surgically resected tissues. Conventional detection methods are invasive and cannot be performed routinely. Therefore, detecting biomarkers on imaging without histopathologic examination would offer significant beneficial. Magnetic resonance imaging (MRI) is one of the most accurate noninvasive methods for diagnosing CRC and can provide additional lesion and functional information through multiple sequence imaging such as T1-weighted image (T1WI), T2-weighted image (T2WI), diffusion weighted imaging (DWI), and dynamic contrast-enhanced (DCE). Colorectal MRI can assess the tumor location and size, infiltration depth, tumour stage, extramural vascular vessels invasion, and circumferential resection margins, thereby facilitating risk stratification [18,19]. Moreover, MRI can be used to evaluate the treatment response and prognosis of patients after neoadjuvant therapy [20].

Radiomics, a noninvasive imaging biomarker, has been used to objectively extract rich quantitative features and establish disease prediction models through machine learning and artificial intelligence (AI) algorithms; many of these radiomics-extracted features are imperceptible to the naked eye of radiologists. Radiomics can effectively provide more comprehensive and accurate information to capture tumour the heterogeneity and has been used to determine tumour types, monitor therapeutic effects, and predict disease prognosis [21–23]. MRI-detected biomarkers in rectal cancer could recognize tumour features and improve diagnostic efficacy [24]. Accordingly, radiomics may assist clinicians in determining CRC risk stratification and provide references for clinical decision-making.

In such a complex oncological scenario, comprehensive management of CRC remains challenging. Traditional tissue biopsy and postoperative pathological examination are invasive, complicated, and time consuming, making it difficult to perform multiple examinations during the disease process. Herein, we attempted to develop and validate a rapid and accurate radiomics predictive models for multiple biological characteristics of CRC based on preoperative clinicopathological data and T2W - MRI features. Additionally, some clinical data were incorporated to establish the corresponding combined models.

2. Methods

2.1. Screening of study participant

The Second Affiliated Hospital Ethics Committee of Fujian Medical University approved this study. This retrospective study enrolled CRC patients who underwent direct surgical resection from May 2022 to August 2023. The exclusion criteria were as follows: (1) absence of preoperative 3.0T MRI scans, (2) no obvious lesions found on MRI, (3) poor image quality, (4) preoperative neoadjuvant therapy, (5) history of previous CRC surgery, (6) missing clinical and pathological data, and (7) presence of other tumors. Finally, 136 CRC patients were included and randomly divided into training cohort and validation cohorts in a 4:1 ratio.

2.2. Clinicopathological data acquisition

Clinicopathological data included age, sex, carcinoembryonic antigen (CEA) and carbohydrate antigen 19 - 9 (CA19 - 9), tumour location, tumour-node-metastasis (TNM) classification, maximum tumour length (MTL), circumferential percentage (CP), differentiation, microsatellite status, and immunohistochemistry results (IHC; P53, Syn, HER2, PNI, VI). Each indicator was divided into groups, and each group was further sub-divided into training and validation cohorts at a 4:1 ratio. All the indicators were divided into positive and negative in binary classifications. The gold standard references were based on the pathological examination results of the postoperative tissue.

2.3. MRI acquisition and tumour segmentation

All patients underwent 3.0 T enhanced MRI scans within 2 weeks before the operation. The T2W-MRI of the CRC patients were retrieved and exported in DICOM format from the communication system (PACS) of the Second Affiliated Hospital of Fujian Medical University. Subsequently, the images and clinical data of the patients were uploaded to the Darwin Research Platform (https://arxiv.org/abs/2009.00908). Fig. 1 shows the radiomics flowchart. The regions of interest (ROI) of entire tumour were performed by a radiologist with 5 years of work experience and verified by a senior gastroenterologist with >10 years of expertise in interpreting colorectal MRIs. They were informed of the location of the tumour but did not know any other clinical information or pathological results. Radiologist manually segmented the ROI along the axis direction of the lesion on each T2W image. They carefully drew the outline to include all imaging information inside the tumour, avoiding the intestinal contents, adipose tissue, blood vessels, and gas in the bowel.

2.4. Feature extraction

After segmentation, the PyRadiomics package was used to extract radiological features based on T2W - MRI. The original features were divided into geometric, intensity, and textural features. Geometric features refer to the three-dimensional shape of the tumour, while texture features describe the patterns, or the second- and high-order spatial distributions of the intensities. Methods for extracting texture features include the gray-level co-occurrence matrix, gray-level run length matrix, gray-level size zone matrix, neighborhood gray-tone difference matrix, and gray-level dependence matrix.

2.5. Feature selection

The Mann–Whitney U statistical test was used to screen the extracted feature data, and only $p < 0.05$ of radiomic features were retained. Subsequently, Spearman's rank correlation coefficient was employed to calculate the correlation among features, focusing on those with high repeatability, while retaining features with a correlation coefficient exceeding 0.9 between any two features. To maintain the optimal ability to depict features, a greedy recursive deletion strategy was adopted to filter features. Minimum redundancy, maximum relevance (mRMR), and least absolute shrinkage and selection operator (LASSO) regression models were used to construct signatures for the discovery dataset. The final model was constructed utilising the LASSO method with 10 - fold cross - validation, and the radiomics score (Rad-score) for each patient was obtained by retaining a linear combination of features weighted by the model coefficients.

2.6. Radiomics signature and clinical signature construction

Eight machine learning models were included to construct the radiomics models: logistic regression (LR), support vector machine (SVM), k-nearest neighbour (KNN), random forest (RF), extra trees (ET), extreme gradient boosting (XGBOOST), light gradient

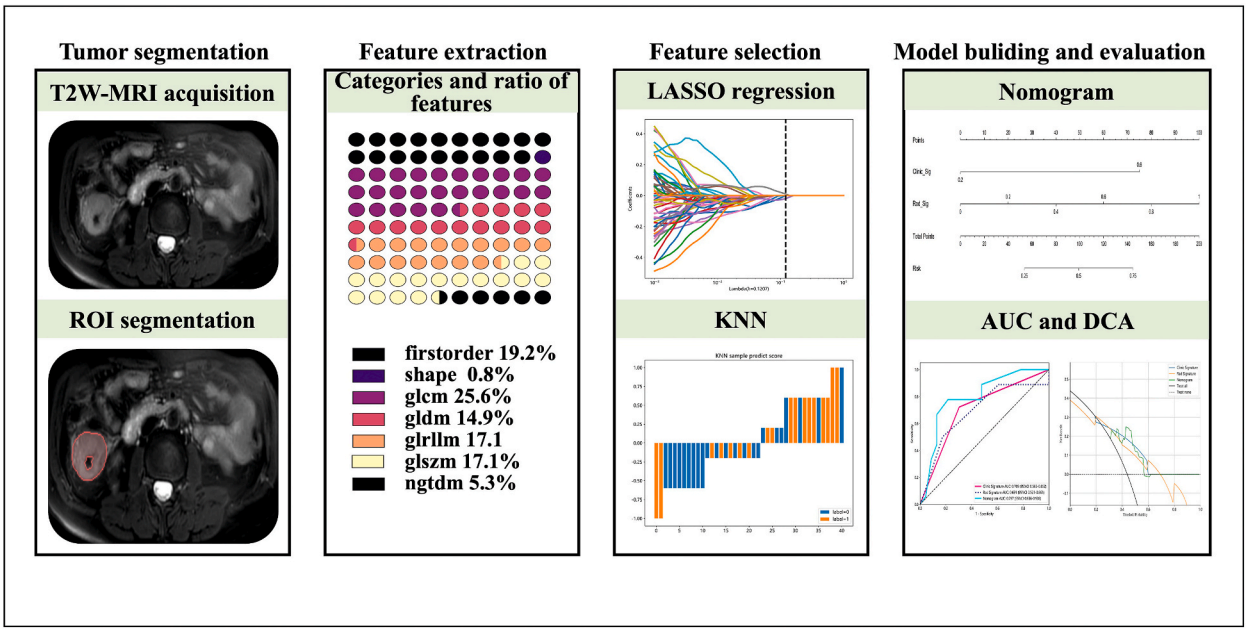

Fig. 1. The radiomics flowchart in this study.

boosting machine (Light GBM), and multilayer perceptron (MLP). Herein, 5-fold cross-validation was employed to derive the final radiomics signature.

Considering the potential predictive value of the clinical parameters, several clinical data (age, sex, CEA, CA19 - 9, TNM classification, MTL, CP, differentiation, tumour location, and microsatellite status) were selected to build the clinical signature. Similarly, we used the abovementioned eight machine learning models described above to construct the clinical signatures. They were integrated into the corresponding T2W - MRI radiomic signature to construct the combined models.

2.7. Statistical analysis

Independent sample t-tests, Mann-Whitney U tests, or chi-square tests were used to select the prospective features from the T2W-MR images or clinical data. Specifically, Independent t-tests were used when the feature data followed a normal distribution, providing means to compare the means of two independent groups. Mann-Whitney U tests were used for non-normally distributed continuous or ordinal data to compare medians between two independent groups. Chi-square tests were applied to categorical data to examine the association between categorical variables. The results, including sample values and t-values, were considered significant at a $P < 0.05$, guiding the selection of features most likely to contribute to predictive accuracy. The diagnostic performance of the models was quantified using receiver operating characteristic (ROC) curves and area under the curve (AUC) along with 95 % confidence intervals (CIs) and compared using the Delong test. Statistical significance was set at $P < 0.05$ was considered statistically significant. A decision curve analysis (DCA) was used to evaluate the net clinical benefits of the constructed model.

3. Results

3.1. Patient characteristics

This study recruited 136 CRC patients (82 males and 54 females; mean age, 63.55 ± 11.55 years; age range 33–88 years); Table 1 lists their baseline characteristics. Based on the pathological results, P53 (94 positives and 42 negatives), Syn (38 positives and 98 negatives), HER2 (58 positives and 78 negatives), PNI (95 positives and 41 negatives), and VI (59 positives and 77 negatives) were selected as the prediction indicators.

Table 1
Patients' baseline characteristics in the training and validation cohorts.

Characteristics	Training cohort	Validation cohort
Age (years), Mean ± SD	63.29 ± 11.76	64.15 ± 11.18
Gender, n (%)		
Male	53 (55.79)	29 (70.73)
Female	42 (44.21)	12 (29.27)
CEA level (ng/ml)	16.78 ± 64.88	8.20 ± 12.56
CA19-9 level (U/ml)	62.93 ± 38.04	59.24 ± 39.41
T stage, n (%)		
T1-2	22 (23.16)	8 (19.51)
T3-4	73 (76.84)	33 (80.49)
N stage, n (%)		
N0	49 (51.58)	26 (63.41)
N1-2	46 (48.42)	15 (36.59)
M stage, n (%)		
M0	85 (89.47)	37 (90.24)
M1	10 (10.53)	4 (9.76)
MLT (cm), Mean ± SD	4.59 ± 1.54	4.53 ± 1.42
CP, n (%)		
< whole circumference	60 (63.16)	24 (58.54)
whole circumference	35 (36.84)	17 (41.46)
Differentiation, n (%)		
Well/Moderate	61 (64.21)	25 (60.98)
Poor	34 (35.79)	16 (39.02)
Tumor location, n (%)		
Colon	26 (27.37)	14 (34.15)
Rectum	69 (72.63)	27 (65.85)
Microsatellite status, n (%)		
MSI	1 (1.05)	3 (7.32)
MSS	94 (98.95)	38 (92.68)

CA199 carbohydrate antigen 199, CEA carcinoembryonic antigen, MLT maximum tumor length, CP circumferential percentage, MSI microsatellite instability, MSS microsatellite stability, SD standard deviation, n number, % percentage.

3.2. Feature extraction process

A total of 1781 radiomic features were extracted from the T2W - MR images; Fig. 2 shows all features and corresponding p-values for each group. After applying Spearman's rank correlation coefficient and greedy recursive deletion strategy to the extracted features and then applying mRMR for the dimension reduction of Syn, PNI, and VI, we retained 61 features for P53, six for Syn, six for HER2, 12 for PNI, and 16 for VI, respectively. For the remaining features, non - zero coefficients were selected to establish the Rad-score using a LASSO logistic regression model. Fig. 3 illustrates the coefficients and mean standard error (MSE) of the 10-fold validation. LASSO logistic analysis showed that P53, Syn, HER2, PNI, and VI had 2, 4, 5, 9, and 9 non-zero coefficient values, respectively (Fig. 4).

3.3. Acquisition of radiomic model

Eight machine learning models (LR, SVM, KNN, RF, ET, XGBOOST, Light GBM, and MLP) were used to construct the radiomics models; Table 2 reveals and their AUCs in each group. Based on the AUC values, KNN performance was the most significant and was selected as the prediction model for the five groups (P53 = 0.663, Syn = 0.683, HER2 = 0.635, PNI = 0.674, and VI = 0.691). For PNI and VI, while RF and ET showed relatively high AUC values in the validation cohorts, these values tended to approach 1 in the training cohorts. Similarly, although XGBOOST achieved an AUC of 0.676 for HER2, it was not selected owing to model overfitting.

3.4. Performance of clinical models and integrated prediction models

Regarding clinical data, only features with $p < 0.05$ in univariate analysis were included in the multivariate logistic analysis. Based on the univariate analysis results, T stage and differentiation were incorporated into the multivariate logistic analysis of P53, Syn, and HER2. Multivariate logistic analysis included the T stage, N stage, and CP of the PNI group and the N and M stages of the VI group. T stage and differentiation were identified as independent risk factors for predicting P53, Syn, and HER2. Moreover, the independent risk factor for PNI and VI was N stage (Table 3). Subsequently, a radiomics nomogram was developed by integrating the corresponding clinical features and Rad-scores of each group (Fig. 5). Accordingly, in the training cohort, the combined model was significantly better than the T2W-MRI models, except for Syn and HER2, based on Delong's tests ($P < 0.05$). Nevertheless, in the validation cohort, except for the Syn group, the combined model was significantly superior to the T2W - MRI model (Table 4; Fig. 6). The clinical radiomics prediction model of PNI performed the best among the five groups, with an AUC value of 0.835 in the validation cohort. The second-best model was VI, with an AUC value of 0.797.

Fig. 7 shows the calibration curves for each group. According to the Hosmer–Lemeshow test results, the clinical radiomics model and clinical models in the training and validation cohorts were consistent with the actual results ($P > 0.05$). Nevertheless, the consistency between the actuality and the prediction models of P53 and HER2 in the validation cohort was poor. In addition, DCA indicated that the prediction model of each group could provide clinical net benefits within a certain range of reasonable threshold probabilities (Fig. 8).

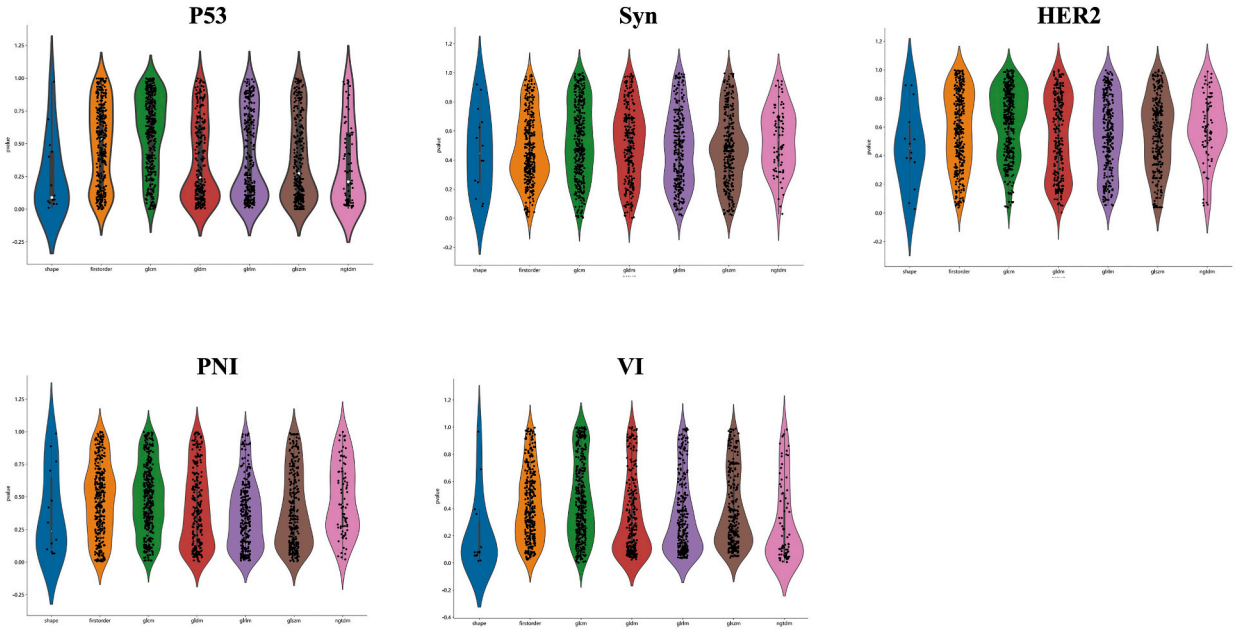

Fig. 2. Statistics of radiomic features for each group.

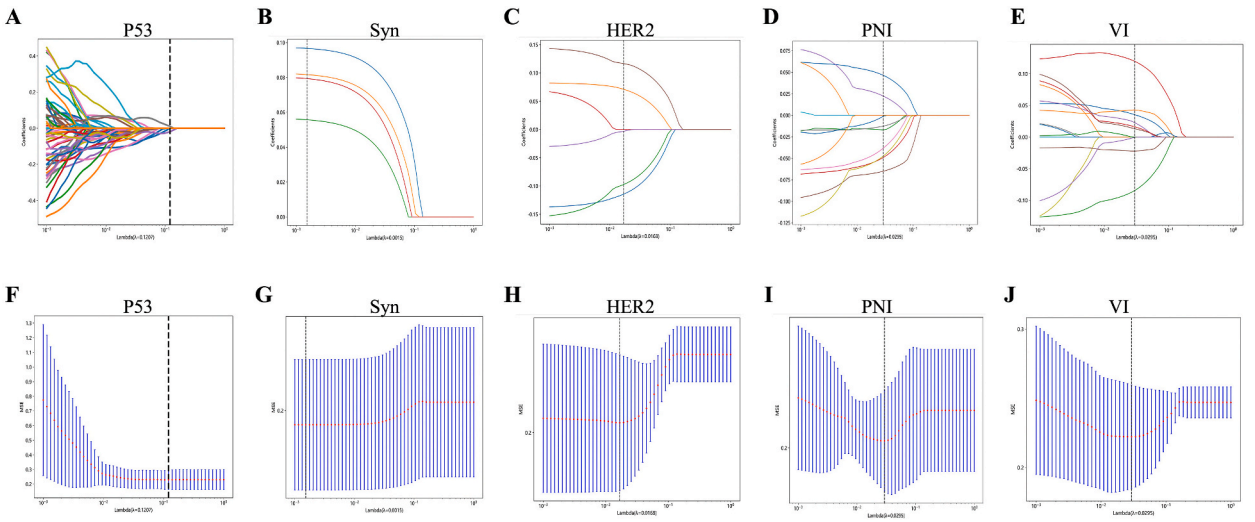

Fig. 3. Coefficients (A–E) and MSE (F–J) of 10-fold cross validation.

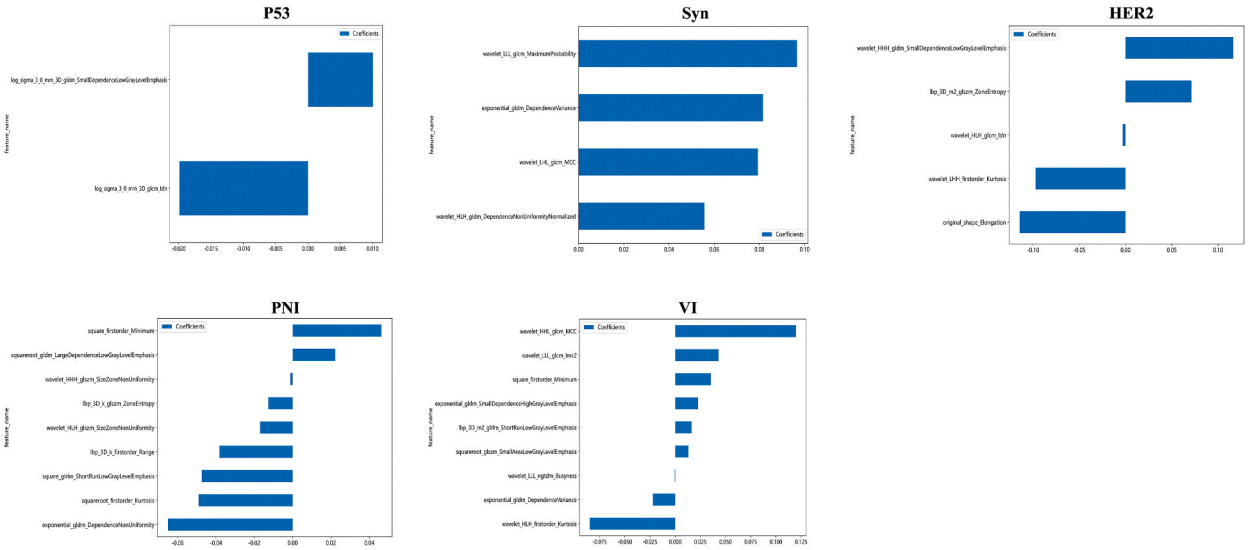

Fig. 4. The histogram of the Rad - score based on the selected features in each group.

Table 2
AUCs of T2W-MRI radiomic models for five predictors in the validation cohorts.

Classifiers	P53	Syn	HER2	PNI	VI
LR	0.619	0.642	0.488	0.477	0.606
SVM	0.400	0.497	0.527	0.671	0.640
KNN	**0.663**	**0.683**	**0.635**	**0.674**	**0.691**
RF	0.705	0.520	0.541	0.750	0.708
ET	0.610	0.577	0.521	0.716	0.714
XGBOOST	0.605	0.550	0.676	0.548	0.626
Light GBM	0.606	0.597	0.477	0.531	0.579
MLP	0.584	0.642	0.510	0.506	0.604

Syn synaptophysin, HER2 human epidermal growth factor receptor 2, PNI perineural invasion, VI vascular invasion, LR logistic regression, SVM support vector machine, KNN k-nearest neighbor, RF random forest, ET extra trees, XGBOOST extreme gradient boosting, Light GBM light gradient boosting machine, MLP multilayer perceptron.

Table 3
Univariate and multivariate logistic regression analysis of clinical data for predicting five biological characteristics.

Predictor	Variables	Univariate logistic	Multivariate logistic		
		p-value	OR	OR 95%CI	p-value
P53	T stage	0.002	1.293	(1.112, 1.502)	0.005
	Differentiation	0.001	0.785	(0.689, 0.893)	0.002
Syn	T stage	0.000	1.295	(1.149, 1.461)	0.002
	Differentiation	0.003	0.821	(0.728, 0.926)	0.007
HER2	T stage	0.000	1.276	(1.120, 1.455)	0.002
	Differentiation	0.001	0.739	(0.647, 0.844)	0.000
PNI	T stage	0.044	–	–	–
	CP	0.029	–	–	–
	N stage	0.000	1.509	(1.339, 1.701)	0.000
LVI	M stage	0.005	–	–	–
	N stage	0.000	1.565	(1.374, 1.784)	0.000

Syn synaptophysin, HER2 human epidermal growth factor receptor 2, PNI perineural invasion, VI vascular invasion, CI Confidence Interval, OR odds ratio.

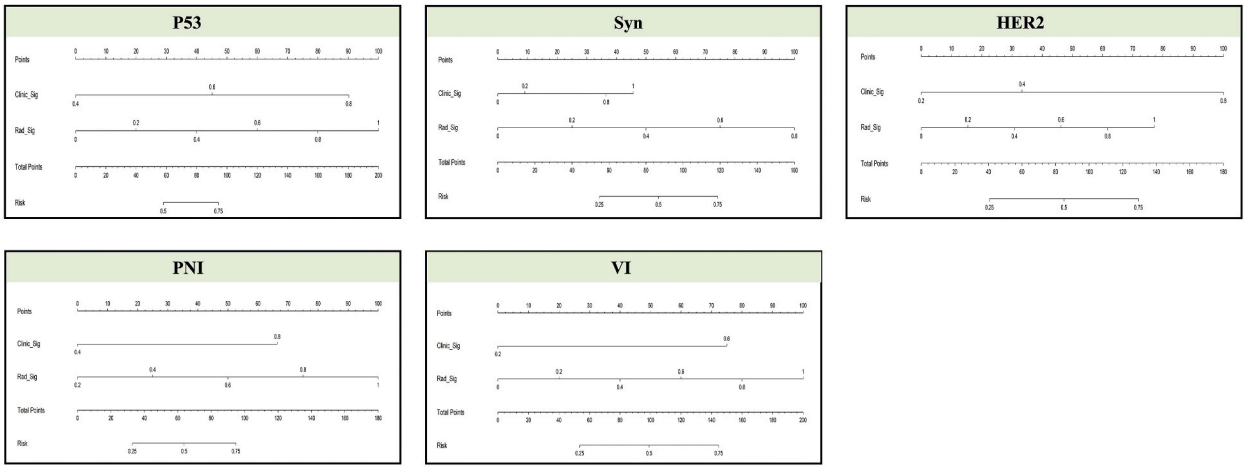

Fig. 5. Nomogram developed for prediction of P53, Syn, HER2, PNI and VI.

4. Discussion

CRC invasiveness has several risk factors: lymph node (LN) metastasis, venous invasion, PNI, P53, HER2, and tumour budding [9, 25–27]. Herein, we developed and validated a clinical radiomics model based on T2W-MRI parameters to preoperatively predict the five biological characteristics (P53, Syn, HER2, PNI, and VI) in CRC patients. To the best of our knowledge, limited studies have simultaneously included multiple biological characteristics to predict CRC prognosis. In our study, we used multiple methods to filter features, reduce dimensionality, and select the optimal machine learning model to construct a radiomics model. Additionally, we utilized a radiomics nomogram to integrate radiomics signature and clinical risk factors. Our results showed that the clinical radiomics model had good diagnostic efficiency and was superior to a single radiomics model and a clinical model in predicting of biological characteristics.

Some studies have used radiomics models and deep learning models to study LN metastasis in CRC [28]. Zhang et al. [29] reported on the diagnostic efficiency of a clinical radiomics model for evaluating of lymphovascular invasion (LVI)/PNI in CRC, based on T2W and diffusion-weighted MRI. The proposed nomogram showed similar performance with an AUC of 0.88 for validation and an AUC of 0.83 for independent test cohorts, compared to our clinical radiomics model which had a PNI-AUC of 0.805 for validation. However, their patient cohorts were differed from ours; additionally, they used LVI/PNI as a combined predictor instead of a separate predictor. In addition, Meng et al. [30] investigated a multiparametric MRI radiomics model of LN metastasis; tumour differentiation; and Ki-67, HER2, and KRAS-2 gene mutation status in rectal cancer. Conversely, our study included both rectal and colon cancer patients as research subjects.

Methods for the preoperative prediction of the biological characteristics of CRC are scarce. Typically, the expression of biological characteristics is evaluated using specimens from invasive biopsies or surgical procedures through IHC. This method, while inexpensive, it is influenced by tissue sample fixation, neoadjuvant chemotherapy, and other subjective factors. Therefore, studies use radiomics has been utilized to develop predictive models owing to its non-i invasiveness, making it suitable for patients who cannot undergo surgery. With the development of radiomics technology, it has become feasible to quickly extract thousands of quantitative

Table 4
Diagnostic efficiency of optimal models for each group.

models	Training cohort					Validation cohort				
	AUC	SEN	SPE	ACC	*p*-value	AUC	SEN	SPE	ACC	*p*-value
P53										
T2W-MRI	0.760	0.524	0.844	0.632	0.004	0.663	0.677	0.600	0.659	0.046
Clinical	0.703	0.429	1.000	0.589		0.698	0.871	0.444	0.756	
Combined	0.821	0.825	0.719	0.789		0.758	0.677	0.800	0.707	
Syn										
T2W-MRI	0.812	0.704	0.809	0.779	0.613	0.683	0.636	0.700	0.683	0.618
Clinical	0.750	0.556	1.000	0.768		0.691	0.455	1.000	0.756	
Combined	0.832	0.704	0.838	0.800		0.739	0.545	0.867	0.780	
HER2										
T2W-MRI	0.876	0.725	0.873	0.811	0.102	0.635	0.389	0.864	0.634	0.048
Clinical	0.726	0.425	1.000	0.716		0.717	0.556	1.000	0.732	
Combined	0.908	0.875	0.855	0.863		0.756	0.889	0.682	0.756	
PNI										
T2W-MRI	0.865	0.812	0.839	0.821	0.013	0.674	0.58	0.889	0.634	0.049
Clinical	0.754	0.734	1.000	0.747		0.756	0.774	1.000	0.756	
Combined	0.930	0.938	0.806	0.895		0.835	0.839	0.700	0.805	
VI										
T2W-MRI	0.760	0.732	0.704	0.716	0.010	0.691	0.500	0.864	0.683	0.043
Clinical	0.758	0.756	1.000	0.758		0.709	0.722	1.000	0.707	
Combined	0.867	0.732	0.852	0.800		0.797	0.778	0.818	0.780	

The models of the five biological characteristics were based on KNN; The p-values were derived from DeLong's test, and they compare the AUCs of the T2W-MRI models with the corresponding combined model. Syn synaptophysin, HER2 human epidermal growth factor receptor 2, PNI perineural invasion, VI vascular invasion.

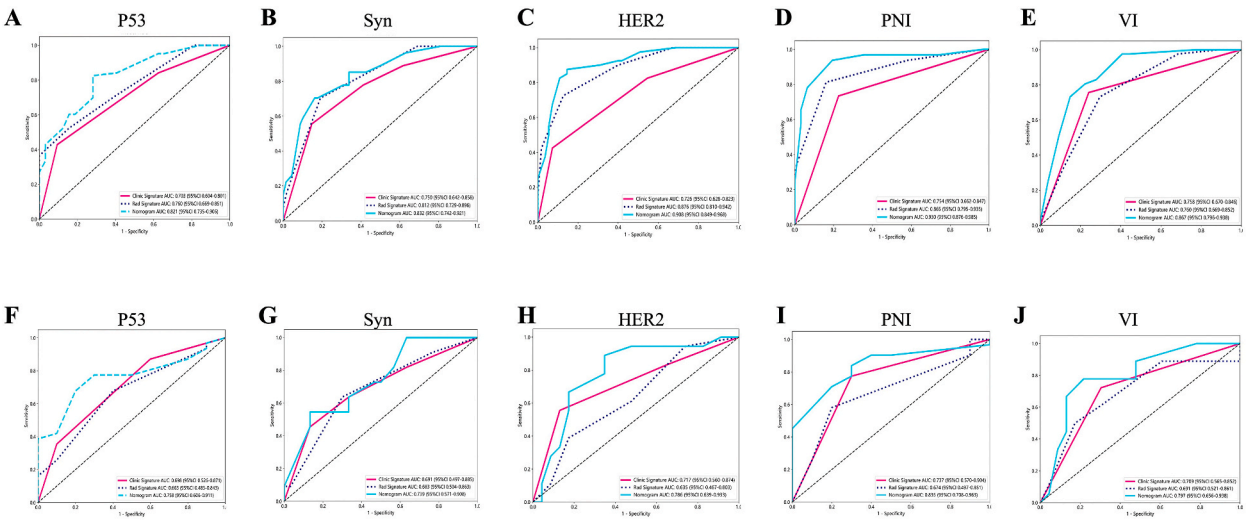

Fig. 6. The AUC curves of the T2W-MRI, clinical and combined models in the training cohort (A–E) and validation cohort (F–J).

features from MRI, computed tomography (CT), ultrasound, and positron emission tomography (PET). Combined with other clinical data (e.g., laboratory results and genomic and proteomic analyses), these features can be used for cancer diagnosis, prognostic evaluation, treatment response prediction, and disease progression monitoring. According to national guidelines, rectal MRI is the preferred method for diagnosing rectal cancer [19]. Compared with CT, MRI undoubtedly has significant advantages in evaluating tumour location, tumour staging, invasion depth, and peripheral soft tissue invasion [18,20,31], which are associated with patient prognosis. Accordingly, we selected T2W-MRI for radiomic features extraction.

Our study initially extracted 1781 features, consistent with recent research records. We utilized mRMR and LASSO regression to generate the best-performing radiomics signatures to the greatest extent and evaluated eight machine learning models for constructing radiomics models; Our empirical results demonstrated that KNN outperformed the other models in terms of accuracy and stability when applied to our dataset. KNN is known for its effectiveness in handling datasets with complex distributions, making it particularly suitable for the heterogeneous and high-dimensional nature of medical imaging data. This algorithm classifies new cases based on a majority vote from the 'k' nearest examples in the feature space, providing a flexible and intuitive approach. This capability is crucial for our study as it allows the model to adapt to the intrinsic variability within radiomic features without making strong parametric

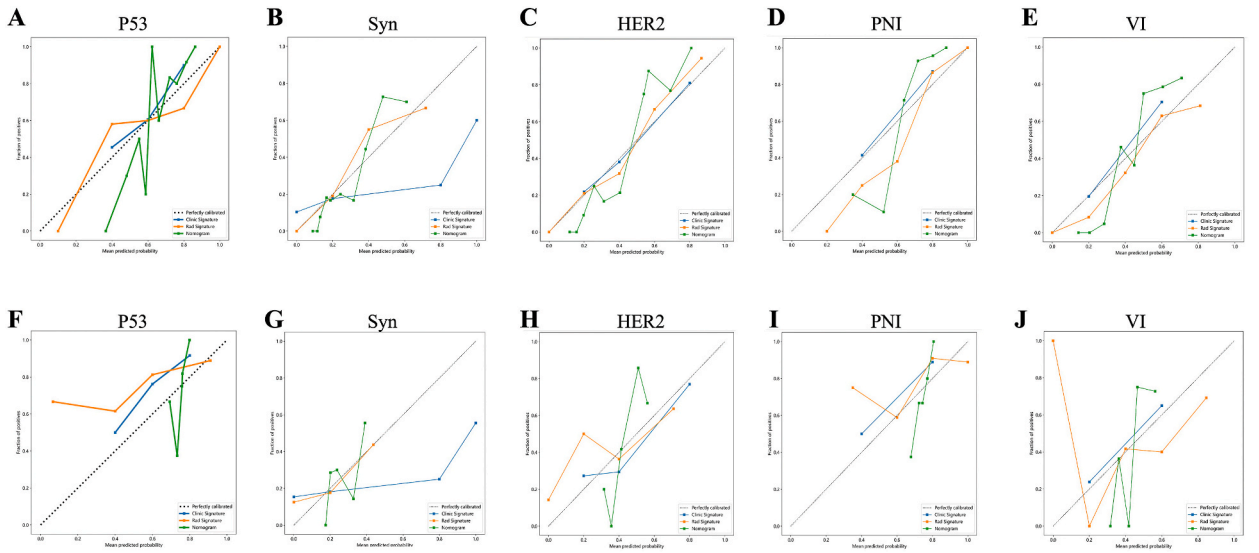

Fig. 7. The calibration curves of the T2W - MRI models, clinical and combined models in the training cohort (A–E) and validation cohort (F–J).

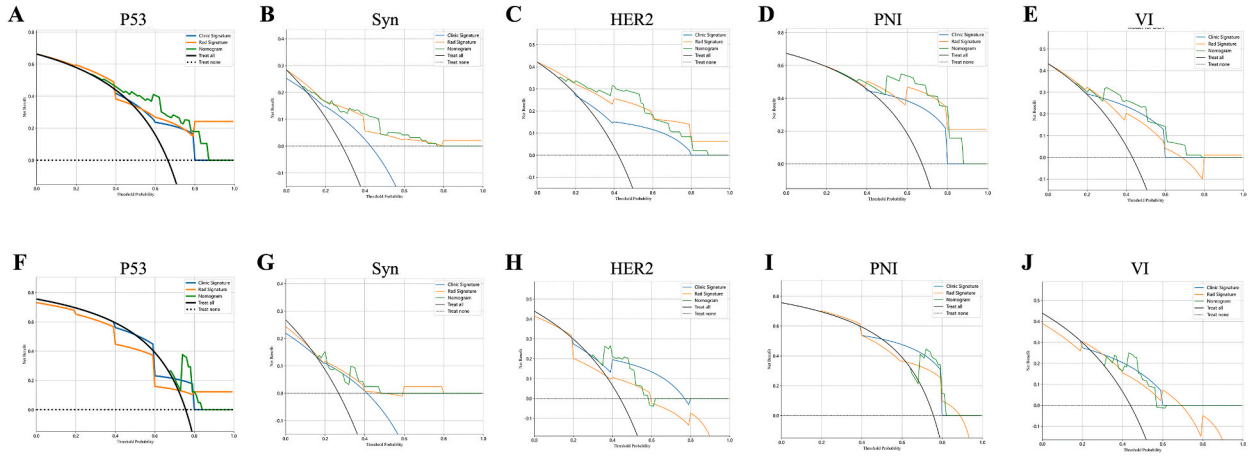

Fig. 8. The DCA curves of the T2W - MRI models, clinical and combined models in the training cohort (A–E) and validation cohort (F–J).

assumptions. Furthermore, the simplicity and non-parametric nature of KNN make it robust against overfitting, especially in scenarios where the relationship between features is not explicitly linear. These characteristics ensure that KNN can capture the nuanced patterns that are often present in detailed medical images, making it an ideal choice for our predictive modeling tasks.

The application of machine learning in CRC has been extended to histopathological examination in order to rapidly identify CRC tissue and its different histopathological growth patterns. A single indicator is usually insufficient to accurately predict patient prognosis accurately. Consequently, we selected multiple biological characteristics of CRC to construct the prediction models. In this study, the AUC values of clinical radiomics model of five biological characteristics all greater than 0.7 in the validation cohort, which indicating that our prediction model has greater advantages in predicting patient prognosis. In terms of clinical utility, the calibration curve and decision curve analysis showed that the clinical radiomics model provided a higher overall net gain.

4.1. Limitation of the proposed work

However, our study had certain limitations. Firstly, this retrospective study was conducted at a single institution, which may have led to some selection bias and drawbacks in external verification. Therefore, external validation using a multicentre database is required to confirm our conclusions. Secondly, we only included the radiomics features of T2W-MRI images; consequently, more MRI sequences must be employed, such as DWI, DCE, and magnetic resonance spectroscopy, must be employed. Thirdly, the relatively small number of included cases limited our use of deep learning to construct predictive models and further inclusion of additional cases and adopt existing deep learning architecture is required to demonstrate greater reliability of our results.

5. Conclusion

Postoperative histopathological examination of excised tissues remains the best choice for diagnosing CRC and identifying its differentiation grade, but it is invasive and requires highly trained laboratory personnel and specialized equipment, with a relatively long time between sampling and the result. Therefore, we used the radiomics features based on T2W-MRI to preoperatively predict P53, Syn, HER2, PNI, and VI expression levels in CRC. The KNN algorithm was used to explore the pathological characteristics of CRC, to identify potential biomarkers, and to construct a practical model with favorable predictive abilities. MRI examination is noninvasive and can be performed multiple times. From this perspective, proposed predictive models could compensate for the limitation of tissue biopsy and quantify the risk of individual patients in terms of tumorigenesis, progression, and metastasis. In addition, our study demonstrated the power of machine learning in cancer research. Machine learning combination with different datasets can provide rapid and reliable techniques to generate knowledge of diseases. In the future, we hope to analyze the omics data and clinical data to predict overall survival, response to chemotherapy, and survival after surgical resection or chemotherapy of CRC. A prompt and accurate predictive model of cancer diagnosis would be useful for optimizing treatment and monitoring.

List of abbreviations

MRI	magnetic resonance imaging
CRC	colorectal cancer
Syn	synaptophysin
IHC	immunohistochemistry
HER2	human epidermal growth factor receptor 2
PNI	perineural invasion
VI	vascular invasion
T2W - MRI	T2-weighted image of MRI
mRMR	minimum redundancy maximum relevance
LASSO	least absolute shrinkage and selection operator regression
ROC	receiver operating characteristic
AUC	area under the curve
KNN	K-nearest neighbor
DCA	decision curve analysis
NED	neuroendocrine differentiation
DWI	diffusion weighted imaging
T1WI	T1 - weighted image
DCE	dynamic contrast - enhanced
T2WI	T2 - weighted image
AI	artificial intelligence
CEA	carcinoembryonic antigen
CA19 - 9	carbohydrate antigen 19 - 9
TNM	tumor location, tumor-node-metastasis
MTL	maximum tumor length
CP	circumferential percentage
ROI	regions of interest
GLCM	gray - level cooccurrence matrix
GLRLM	gray - level run length matrix
GLSZM	gray - level size zone matrix
NGTDM	neighborhood gray - tone difference matrix
GLDM	gray - level dependence matri
Rad - score	radiomics score
LR	logistic regression
SVM	support vector machine
RF	random forest,
ET	extra trees
XGBOOST	extreme gradient boosting
Light GBM	light gradient boosting machine
MLP	multilayer perceptron
CIs	confidence intervals
MSE	mean standard error
LN	lymph node
LVI	lymphovascular invasion
CT	Computed Tomography
PET	positron emission tomography

CRediT authorship contribution statement

Qiao-yi Huang: Writing – original draft. **Hui-da Zheng:** Writing – review & editing. **Bin Xiong:** Methodology. **Qi-ming Huang:** Data curation. **Kai Ye:** Validation. **Shu Lin:** Supervision. **Jian-hua Xu:** Project administration.

Ethics approval and consent to participate

This study was approved by the Second Affiliated Hospital of Fujian Medical University provided ethical approval for this study (approval#: 2023-430). All persons gave their informed consent prior to their inclusion in the study.

Consent for publication

Not applicable.

Data availability statement

The data that support the findings of this study are included in article/supp. material/referenced in article.

Funding

This work was supported by a grant from Malignant Tumor Clinical Medicine Research Cenrer (No.2020N090s), the Youth Research Project of Fujian Provincial Health Commission (No.2022QNA067), and Fujian Province Science and Technology Innovation Joint Fund Project (Grant number 2023Y9254).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Home for Researchers editorial team (www.home-for-researchers.com) for language editing service.

References

[1] H. Sung, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 71 (3) (2021) 209–249.
[2] M. Arnold, et al., Global patterns and trends in colorectal cancer incidence and mortality, Gut 66 (4) (2017) 683–691.
[3] H. Zhou, et al., Liquid biopsy at the frontier of detection, prognosis and progression monitoring in colorectal cancer, Mol. Cancer 21 (1) (2022) 86.
[4] K.M. Kim, et al., Clinical significance of p53 protein expression and TP53 variation status in colorectal cancer, BMC Cancer 22 (1) (2022) 940.
[5] D. Fang, H. Hu, K. Zhao, A. Xu, C. Yu, Y. Zhu, N. Yu, B. Yao, S. Tang, X. Wu, Y. Mei, MLF2 negatively regulates P53 and promotes colorectal carcinogenesis, Adv. Sci. 10 (26) (2023 Sep) e2303336.
[6] Y. Hou, X. Zhang, H. Yao, L. Hou, Q. Zhang, E. Tao, X. Zhu, S. Jiang, Y. Ren, X. Hong, S. Lu, X. Leng, Y. Xie, Y. Gao, Y. Liang, T. Zhong, B. Long, J.Y. Fang, X. Meng, METTL14 modulates glycolysis to inhibit colorectal tumorigenesis in p53-wild-type cells, EMBO Rep. 24 (4) (2023 Apr 5) e56325.
[7] X. Liu, Y. Liu, Z. Liu, C. Lin, F. Meng, L. Xu, X. Zhang, C. Zhang, P. Zhang, S. Gong, N. Wu, Z. Ren, J. Song, Y. Zhang, CircMYH9 drives colorectal cancer growth by regulating serine metabolism and redox homeostasis in a p53-dependent manner, Mol. Cancer 20 (1) (2021 Sep 8) 114.
[8] T. Gagliano, E. Kerschbamer, U. Baccarani, M. Minisini, E. Di Giorgio, E. Dalla, C.X. Weichenberger, V. Cherchi, G. Terrosu, C. Brancolini, Changes in chromatin accessibility and transcriptional landscape induced by HDAC inhibitors in TP53 mutated patient-derived colon cancer organoids, Biomed. Pharmacother. 173 (2024 Apr) 116374.
[9] C.J. Shen, R.H. Chan, B.W. Lin, N.C. Li, Y.H. Huang, W.C. Chang, B.K. Chen, Oleic acid-induced metastasis of KRAS/p53-mutant colorectal cancer relies on concurrent KRAS activation and IL-8 expression bypassing EGFR activation, Theranostics 13 (13) (2023 Aug 21) 4650–4666.
[10] B. Kleist, M. Poetsch, Neuroendocrine differentiation: the mysterious fellow of colorectal cancer, World J. Gastroenterol. 21 (41) (2015) 11740–11747.
[11] M. Fassan, et al., Synaptophysin expression in (V600EBRAF-)mutated advanced colorectal cancers identifies a new subgroup of tumours with worse prognosis, Eur. J. Cancer 146 (2021) 145–154.
[12] Y. Liu, et al., Neuroendocrine differentiation is predictive of poor survival in patients with stage II colorectal cancer, Oncol. Lett. 13 (4) (2017) 2230–2236.
[13] V.A. Afrăsânie, et al., KRAS, NRAS, BRAF, HER2 and microsatellite instability in metastatic colorectal cancer - practical implications for the clinician, Radiol. Oncol. 53 (3) (2019) 265–274.
[14] J.S. Pyo, G. Kang, K. Park, Clinicopathological significance and diagnostic accuracy of HER2 immunohistochemistry in colorectal cancer: a meta-analysis, Int. J. Biol. Markers 31 (4) (2016) e389–e394.
[15] H. Nozawa, et al., Obstruction is associated with perineural invasion in T3/T4 colon cancer, Colorectal Dis. 21 (8) (2019) 917–924.
[16] J. Betge, et al., Intramural and extramural vascular invasion in colorectal cancer: prognostic significance and quality of pathology reporting, Cancer 118 (3) (2012) 628–638.
[17] T. Fujii, et al., Vascular invasion, but not lymphatic invasion, of the primary tumor is a strong prognostic factor in patients with colorectal cancer, Anticancer Res. 34 (6) (2014) 3147–3151.
[18] X. Jiang, H. Zhao, O.L. Saldanha, S. Nebelung, C. Kuhl, I. Amygdalos, S.A. Lang, X. Wu, X. Meng, D. Truhn, J.N. Kather, J. Ke, An MRI deep learning model predicts outcome in rectal cancer, Radiology 307 (5) (2023 Jun) e222223.
[19] P.P. Wang, C.L. Deng, B. Wu, Magnetic resonance imaging-based artificial intelligence model in rectal cancer, World J. Gastroenterol. 27 (18) (2021) 2122–2130.
[20] J. Shin, N. Seo, S.E. Baek, N.H. Son, J.S. Lim, N.K. Kim, W.S. Koom, S. Kim, MRI radiomics model predicts pathologic complete response of rectal cancer following chemoradiotherapy, Radiology 303 (2) (2022 May) 351–358.
[21] D. Caruso, et al., Radiomics in oncology, Part 1: technical principles and gastrointestinal application in CT and MRI, Cancers 13 (11) (2021).
[22] D. Caruso, et al., Radiomics in oncology, Part 2: thoracic, genito-urinary, breast, neurological, hematologic and musculoskeletal applications, Cancers 13 (11) (2021).
[23] F. Botta, et al., Association of a CT-based clinical and radiomics score of non-small cell lung cancer (NSCLC) with lymph node status and overall survival, Cancers 12 (6) (2020).

[24] Y. Sun, et al., Radiomic features of pretreatment MRI could identify T stage in patients with rectal cancer: preliminary findings, J. Magn. Reson. Imag. 48 (3) (2018) 615–621.
[25] C.F. Rönnow, et al., Lymphovascular infiltration, not depth of invasion, is the critical risk factor of metastases in early colorectal cancer: retrospective population-based cohort study on prospectively collected data, including validation, Ann. Surg. 275 (1) (2022) e148–e154.
[26] M. Swets, et al., Are pathological high-risk features in locally advanced rectal cancer a useful selection tool for adjuvant chemotherapy? Eur. J. Cancer 89 (2018) 1–8.
[27] W. Huang, et al., HER2 positivity as a biomarker for poor prognosis and unresponsiveness to anti-EGFR therapy in colorectal cancer, J. Cancer Res. Clin. Oncol. 148 (4) (2022) 993–1002.
[28] S. Bedrikovetski, et al., Artificial intelligence for pre-operative lymph node staging in colorectal cancer: a systematic review and meta-analysis, BMC Cancer 21 (1) (2021) 1058.
[29] K. Zhang, et al., A clinical-radiomics model incorporating T2-weighted and diffusion-weighted magnetic resonance images predicts the existence of lymphovascular invasion/perineural invasion in patients with colorectal cancer, Med. Phys. 48 (9) (2021) 4872–4882.
[30] X. Meng, et al., Preoperative radiomic signature based on multiparametric magnetic resonance imaging for noninvasive evaluation of biological characteristics in rectal cancer, Eur. Radiol. 29 (6) (2019) 3200–3209.
[31] A.C. Lord, N. D'Souza, A. Shaw, Z. Rokan, B. Moran, M. Abulafi, S. Rasheed, A. Chandramohan, A. Corr, I. Chau, G. Brown, MRI-diagnosed tumor deposits and EMVI status have superior prognostic accuracy to current clinical TNM staging in rectal cancer, Ann. Surg. 276 (2) (2022 Aug 1) 334–344.