Research Square

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

Preoperative prediction of microsatellite instability in rectal cancer using multiparametric MRIderived intratumoral heterogeneity and exploration of biological mechanisms: a multicenter study

LiXin LI
Lanzhou University Second Hospital
Duo Yang
Lanzhou University Second Hospital
Yuxin Yan
Lanzhou University Second Hospital
Wanjun Hu
Lanzhou University Second Hospital
Yingying Jiang
Gansu University of Traditional Chinese Medicine
Peng Liu
Gansu Provincial Cancer Hospital Gansu Provincial Academic Instiute for Medical Research
Yawei Shi
Department of Pathology, Lanzhou University Second Hospital
Yuping Bai
309131762@qq.com

Lanzhou University Second Hospital

Research Article

Keywords: Rectal cancer, Microsatellite instability, Radiomics, Intratumoral heterogeneity, Magnetic resonance imaging

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10721594/v1

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

Additional Declarations: No competing interests reported.

Preoperative prediction of microsatellite instability in rectal cancer using multiparametric MRI-derived intratumoral heterogeneity and exploration of biological mechanisms: a multicenter study

ABSTRACT

Objective To develop a multi-parametric MRI-based intratumoral heterogeneity (ITH) scoring model combined with clinical features for preoperative prediction of microsatellite instability (MSI) in rectal cancer, and to explore MSI-related molecular features.

Methods A retrospective multicenter study of 464 rectal cancer patients (training n=198, internal validation n=110, external validation n=156). Three-dimensional ITH scores were derived from T2WI and DWI. Six machine learning algorithms were used to build clinical, ITH, and combined models, evaluated by AUC, calibration curves, and SHAP analysis. TCGA transcriptomic data and immunohistochemistry (n=10) were used to explore MSI-associated molecular features.

Results Age, tumor length, location, and N stage were independent MSI predictors. The combined ExtraTrees model achieved AUCs of 0.865, 0.839, and 0.785 across three cohorts, outperforming single models. PPI network analysis identified AKT1 as the hub gene. MSS tumors showed significantly higher AKT1 expression than MSI-H tumors (H-score: 143.60±26.14 vs. 76.74±38.43, P=0.012).

Conclusions The ITH-clinical combined model enables effective non-invasive preoperative MSI prediction, with ITH providing incremental predictive value. Differential AKT1 expression offers preliminary molecular evidence for understanding MSI in rectal cancer.

Keywords Rectal cancer; Microsatellite instability; Radiomics; Intratumoral heterogeneity; Magnetic resonance imaging

Introduction

Colorectal cancer (CRC) remains a leading cause of cancer-related morbidity and mortality worldwide, accounting for approximately 10% and 15% of global new cases and deaths, respectively, and representing a major public health challenge[1,2]. Microsatellite instability (MSI), a key pathogenic mechanism of CRC, occurs in approximately 15% of cases [3,4]. Patients with MSI-H CRC typically derive significant clinical benefit from immune

checkpoint inhibitors, whereas they exhibit poor response to 5-fluorouracil (5-FU)-based adjuvant chemotherapy alone [5-7]. Therefore, accurate, non-invasive preoperative identification of MSI status is of great clinical significance for guiding individualized treatment strategies and optimizing patient outcomes.

Currently, the clinical gold standard for MSI detection relies on invasive laboratory analyses of postoperative or endoscopic biopsy specimens, including polymerase chain reaction (PCR) and immunohistochemistry (IHC). Beyond their invasiveness, these approaches cannot comprehensively assess phenotypic variations across different intratumoral regions — that is, intratumoral heterogeneity (ITH). Growing evidence indicates that CRC harbors substantial ITH, with MSI status and molecular profiles potentially varying across different regions within the same tumor; consequently, single-biopsy sampling may lead to over- or underestimation of MSI status [8,9]. There is therefore an urgent need for a non-invasive preoperative approach that can fully capture ITH to more accurately assess MSI status in rectal cancer.

ITH is a critical driver of tumor evolution, immune evasion, and therapeutic resistance [10,11], and its quantitative characterization is of considerable importance for understanding tumor biological behavior and predicting treatment response. In recent years, imaging-based ITH assessment has attracted increasing attention. By segmenting tumors into multiple subregions, extracting radiomic features from each subregion, and quantifying inter-regional differences, these approaches enable non-invasive characterization of phenotypic diversity across intratumoral functional subregions [12,13]. Radiomics extracts high-throughput quantitative features from routine medical images, converting visually imperceptible information into mineable numerical data and offering a new avenue for non-invasive tumor phenotyping [14,15]. Radiogenomics seeks to systematically establish associations between imaging features and genomic data to uncover the molecular basis underlying imaging phenotypes, thereby enhancing model interpretability[16]. To date, radiogenomic models based on CT or MRI have demonstrated considerable potential in CRC. Zhang et al. [17] developed an immunophenotype-guided radiomics model using CT to predict pathological complete response to neoadjuvant anti-PD-1 therapy in stage III-IV d-MMR/MSI-H CRC, achieving an AUC of 0.904 in an independent validation set by classifying tumors into immune-hot and immune-cold subtypes via unsupervised clustering. However, existing studies share two main

limitations: (1) most models employ single-center, whole-tumor mean features, treating tumors as homogeneous entities and failing to adequately capture phenotypic differences among intratumoral functional subregions; and (2) models generally lack clinical interpretability, with insufficient exploration of the biological mechanisms underlying key imaging features, thereby limiting their clinical translation. At present, integrating imaging features with biological mechanistic exploration for constructing and interpreting non-invasive MSI prediction models in rectal cancer remains in its preliminary stages.

Accordingly, the present study introduces an ITH scoring approach to quantitatively characterize intratumoral functional subregions using multi-parametric MRI, and develops and validates a preoperative prediction model for MSI status in rectal cancer. Furthermore, we explore the potential biological mechanisms underlying key imaging features at the molecular level, and employ SHAP (SHapley Additive exPlanations) analysis to quantify the clinical decision contributions of key features, aiming to provide an objective basis for non-invasive preoperative MSI assessment and individualized treatment decision-making.

1 Materials and Methods

1.1 Study Population

This retrospective multicenter study was approved by the institutional review boards of all participating centers, with a waiver of informed consent. Patients were recruited from three tertiary hospitals: Lanzhou University Second Hospital (Center A), Gansu Provincial Cancer Hospital (Center B), and Yunnan Provincial Cancer Hospital (Center C). A total of 308 patients pathologically confirmed with rectal cancer at Lanzhou University Second Hospital between January 2020 and November 2024 (38 MSI-positive) were allocated to the training and internal validation sets. Additionally, 156 patients (23 MSI-positive) enrolled from Gansu Provincial Cancer Hospital and Yunnan Provincial Cancer Hospital between January 2022 and November 2024 constituted the external validation set.

Inclusion criteria were as follows: (1) pathologically confirmed primary rectal adenocarcinoma; (2) preoperative multi-parametric MRI examination; (3) postoperative MSI status determined by immunohistochemical assessment of four mismatch repair (MMR)

proteins (MLH1, MSH2, MSH6, PMS2), with deficiency in any protein defined as mismatch repair deficient (dMMR, MSI-H) and intact expression of all four proteins defined as proficient mismatch repair (pMMR, MSS); and (4) no prior neoadjuvant chemoradiotherapy, targeted therapy, or immunotherapy. Exclusion criteria included: (1) contraindications to MRI; (2) poor MRI image quality with inadequate tumor delineation precluding accurate region-of-interest (ROI) delineation; and (3) missing key clinicopathological data (e.g., staging, laboratory results). The patient enrollment process is summarized in Figure 1.

Clinical variables collected included age, sex, tumor location, tumor length, carcinoembryonic antigen (CEA), CA19-9, CA125, MRI-based T stage (MRI-T stage), and N stage (MRI-N stage). CEA > 5 ng/mL was defined as elevated. MSI status was determined by IHC assessment of the four MMR proteins: loss of nuclear staining for one or more proteins was classified as the MSI group, whereas intact expression of all proteins was classified as the MSS group.

1.2 MRI Acquisition and Preprocessing

All MRI examinations were performed using 3.0T scanners. Patients were not required to undergo bowel preparation; however, an intramuscular injection of 20 mg anisodamine was administered 30 minutes before scanning to reduce bowel peristalsis artifacts, and patients were instructed to empty the rectum. The key sequences analyzed included axial T2-weighted imaging (T2WI) and axial diffusion-weighted imaging (DWI). Detailed scanning parameters are provided in Supplementary Table 1. To mitigate inter-center and inter-scanner variability, all raw DICOM images underwent standardized preprocessing prior to feature extraction, consisting of: (1) bias field correction; (2) resampling to isotropic voxels ($1 \times 1 \times 1$ mm³); and (3) Z-score normalization of image intensity values.

1.3 Tumor Segmentation

Two radiologists with over five years of experience in gastrointestinal tumor imaging, blinded to MSI status, manually delineated the three-dimensional volumes of interest (VOIs) on T2WI slice by slice using ITK-SNAP software (www.itksnap.org), carefully avoiding luminal contents and surrounding perirectal fat. The T2WI VOIs were then rigidly registered to the DWI sequence. Segmentation results were cross-checked by both readers, with

discrepancies resolved through consensus or adjudication by a senior radiologist. Radiomic features were extracted independently from both readers' segmentations, and inter-observer reproducibility was assessed using the intraclass correlation coefficient (ICC), with only features achieving ICC > 0.8 retained for subsequent analysis.

1.4 ITH Score Calculation

In this study, pixel-wise clustering was performed within the region of interest based on the lesions of rectal cancer patients, and ITH scores were calculated for quantitative assessment. This analytical approach was referenced from previous literature[18,19]; this study improved the method and applied it to quantifying ITH from multiparametric MRI scans of rectal cancer patients, aiming to achieve preoperative non-invasive prediction of MSI status. The ITH score calculation steps were as follows: (a) Based on all voxels within the three-dimensional tumor VOI from both T2WI and DWI sequences, a 3 × 3 × 3 mm three-dimensional sliding window was applied, and local three-dimensional radiomics features (including 19 first-order statistical features and 75 three-dimensional texture features) were extracted using PyRadiomics (version 3.1). All features were Z-score standardized, and images were uniformly resampled to 1 × 1 × 1 mm^3 isotropic voxels to eliminate slice thickness variation. (b) K-means clustering was employed to partition the tumor into subregions sharing analogous radiomic feature profiles. For each tumor, the optimal cluster number was determined jointly using the elbow method and the silhouette coefficient – namely, the K value that coincided with the elbow point of the inertia curve and simultaneously maximized the silhouette coefficient. The search range for K was predefined as 2 to 8. (c) The ITH score was calculated based on the number of connected components (n_i) and the maximum area of connected components (S_i,max) for each cluster.

$$\text{ITH score} = 1 - \frac{1}{S_{\text{total}}} \sum_{i=1}^{V} \frac{S_{i,\max}}{n_i},$$

where V is the number of clusters, S_{total} is the total lesion area, n_i is the number of connected regions, and $S_{i,\max}$ is the maximum area of a single connected region within the i-th cluster. The ITH score ranges from 0 to 1. A higher ITH score indicates greater spatial dispersion of imaging feature values within the tumor, reflecting greater intratumoral heterogeneity. To characterize intratumoral heterogeneity across multiple dimensions, the

following derived features were calculated based on the ITH scores of each cluster: (1) ITH mean: the mean of ITH scores across all clusters; (2) ITH minimum: the minimum of ITH scores across all clusters; (3) ITH product: the product of ITH scores across all clusters; (4) ITH absolute difference: the absolute difference between the maximum and minimum cluster ITH scores; (5) ITH signature: the first principal component score obtained through principal component analysis (PCA) dimensionality reduction of the above four derived features together with the global ITH score, serving as a comprehensive heterogeneity label aimed at integrating multi-dimensional ITH information into a single metric. Ultimately, the ITH-derived features along with T2WI ITH score and DWI ITH score were incorporated into the ITH feature set, with stable features retained after ICC screening for model construction. The study workflow is illustrated in Figure 2.

1.5 Data Balancing

Given the relatively low proportion of MSI-positive cases in this study, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set to address class imbalance and improve model performance. SMOTE generates synthetic minority-class samples by interpolating between each original minority sample and its k-nearest neighbors (k = 5). This technique was applied exclusively to the training set and not to the validation sets, in order to reflect the real-world prevalence distribution of MSI and MSS and to prevent label leakage. Following SMOTE, the ratio of MSI to MSS samples in the training set was adjusted to 1:1 (from an original MSI-positive proportion of approximately 19%).

1.6 Feature Selection and Predictive Model Construction

Inter-observer reproducibility of ITH score calculations was first assessed using ICC, with unstable features (ICC < 0.80) excluded and only reproducible features retained for the ITH feature set. Six machine learning algorithms were employed to construct models: logistic regression (LR), Gaussian naive Bayes (GNB), support vector machine (SVM), extremely randomized trees (Extra Trees, ET), random forest (RF), and gradient boosting decision tree (GBDT). Simultaneously, univariate and multivariate logistic regression analyses were performed on clinical-pathological variables to identify predictors of MSI status for constructing the clinical model. Three models were ultimately developed: (a) a clinical model,

(b) an ITH model, and (c) a combined clinical-ITH model.

Model performance was evaluated on the independent internal validation and external validation sets. The primary evaluation metrics included the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. Calibration curves were plotted to assess model calibration, and decision curve analysis (DCA) was performed to evaluate the clinical net benefit of each model.

1.7 Model Interpretability Analysis

To enhance model interpretability, SHAP (SHapley Additive exPlanations) values were computed to quantify the contribution of each feature to individual model predictions. SHAP analysis not only identifies the most important features for MSI prediction but also reveals the relative weights of the ITH score and its multi-dimensional components in the prediction process, thereby enhancing the biological interpretability of the model and providing a theoretical basis for the clinical relevance of ITH scoring.

1.8 Bioinformatics Analysis

RNA sequencing count matrices and corresponding clinical annotation data for colorectal adenocarcinoma (COAD and READ) were downloaded from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). Samples with missing MSI status annotations were excluded. MSI-H and MSI-L samples were combined into the MSI group for comparison with MSS samples. Differential gene expression analysis was performed using the R package DESeq2 (version 1.38.3), with the cancer type (COAD/READ) included as a covariate in the statistical model. Lowly expressed genes were filtered out, retaining genes with a total count greater than 10 across all samples. Differentially expressed genes (DEGs) were defined using thresholds of $|\log_2 \text{fold change}| > 1$ and adjusted P-value (padj) < 0.05. The expression matrix was normalized using variance-stabilizing transformation (VST), followed by row-wise Z-score standardization for heatmap visualization. All significant DEGs were imported into the STRING database (https://string-db.org/, version 11.5) with the species set to *Homo sapiens* and the interaction confidence threshold set to high confidence (combined score ≥ 0.7). Interactions derived from experimental validation, database annotations, and co-expression were retained to construct the protein-protein interaction (PPI) network. The PPI

network data were imported into Cytoscape software (version 3.9.1) for visualization and topological analysis. The Maximal Clique Centrality (MCC) algorithm implemented in the cytoHubba plugin (version 1.0.0), combined with node degree, betweenness centrality, and closeness centrality, was used to comprehensively evaluate the centrality of each node within the network to identify key hub genes.

1.9 Experimental Validation

To validate the bioinformatics findings at the protein level, five MSI-H and five MSS rectal cancer formalin-fixed paraffin-embedded (FFPE) tissue specimens were randomly selected from the training cohort. All patients had received no neoadjuvant therapy prior to surgery, and there were no significant differences between the two groups in age, sex, tumor location, or TNM stage (all $P > 0.05$). Immunohistochemistry was performed to detect AKT1 protein expression. FFPE sections (4 μm) were deparaffinized and rehydrated, followed by antigen retrieval using citrate buffer (pH 6.0) under high pressure for 30 minutes. Endogenous peroxidase activity was blocked by incubation with 3% hydrogen peroxide for 25 minutes at room temperature in the dark. After PBS washing, sections were blocked with 3% bovine serum albumin (BSA) for 30 minutes at room temperature. Sections were incubated overnight at 4°C with a rabbit anti-human AKT1 monoclonal antibody (catalog number: GB15689, Servicebio, dilution 1:500). On the following day, sections were incubated with horseradish peroxidase-conjugated ready-to-use S-vision polymeric secondary antibody (goat anti-rabbit, catalog number: G1302, Servicebio) for 50 minutes at room temperature. DAB chromogenic development was performed under microscopic control, followed by hematoxylin counterstaining, dehydration, clearing, and mounting. A known AKT1-positive breast cancer tissue section was used as the positive control, and PBS replacing the primary antibody was used as the negative control.Two pathologists blinded to the clinical and MSI status independently performed semiquantitative scoring. For each slide, five randomly selected high-power fields (200×) were evaluated for staining intensity (0–3) and the percentage of positive cells (0–100%). The H-score was calculated as the sum of intensity × percentage (range, 0–300). Inter-observer agreement was assessed using ICC. Between-group comparisons of H-scores were performed using the independent-samples t-test or the Mann-Whitney U test, with

P < 0.05 considered statistically significant.

1.10 Statistical Analysis

Clinical data were analyzed using SPSS version 27.0. Normally distributed continuous variables were presented as mean ± standard deviation and compared using the t-test. Non-normally distributed continuous variables were presented as median values and compared using the Mann-Whitney U test. Categorical variables were presented as frequencies and compared using the chi-square test or Fisher's exact test. A two-sided P-value < 0.05 was considered statistically significant.

2 Results

2.1 Clinical Characteristics

A total of 464 patients with pathologically confirmed rectal cancer were enrolled in this study. The modeling cohort comprised 308 patients from Center A, randomly stratified into a training set (n = 198) and an internal validation set (n = 110), while 156 patients from Centers B and C constituted the independent external validation set. The median patient age was 61 years (interquartile range: 55.00–69.00), with 281 males (60.11%) and 183 females (39.44%). The overall prevalence of MSI was 13.15% (61/464). Regarding clinicopathological characteristics, the distributions of T stage and N stage differed significantly among the three sets (all P < 0.05), whereas no significant differences were observed for age, sex, tumor length, tumor location, M stage, MSI status, or serum levels of CA19-9, CA125, and CEA (all P > 0.05). Detailed clinicopathological characteristics are presented in Table 1.

2.2 Clinical Risk Factors

Univariate logistic regression analysis revealed that age (OR = 0.94, 95% CI: 0.91–0.97, P = 0.01), tumor length (OR = 1.02, 95% CI: 1.01–1.03, P = 0.02), tumor location (lower rectum vs. upper rectum as reference: OR = 0.08, 95% CI: 0.01–0.63, P = 0.02), and N stage (N1 vs. N0 as reference: OR = 0.17, 95% CI: 0.04–0.71, P = 0.015) were significantly associated with MSI status. Sex, T stage, and serum levels of CA19-9, CA125, and CEA showed no significant associations (all P > 0.05). Variables with P < 0.10 in the univariate analysis were entered into a multivariate logistic regression model, which identified age (OR = 0.94, 95% CI: 0.91–0.97,

P < 0.01), tumor length (OR = 1.02, 95% CI: 1.01–1.04, P = 0.05), tumor location (lower rectum vs. upper rectum: OR = 0.11, 95% CI: 0.01–0.87, P = 0.04), and N stage (N1 vs. N0: OR = 0.21, 95% CI: 0.05–0.93, P = 0.039) as independent predictors of MSI status. These four variables were therefore incorporated into the clinical model (Table 2).

2.3 Development and Validation of Predictive Models

Six machine learning algorithms — logistic regression (LR), Gaussian naive Bayes (GNB), support vector machine (SVM), random forest (RF), gradient boosting decision tree (GBDT), and extremely randomized trees (ExtraTrees, ET) — were trained using the combined clinical-ITH feature set and evaluated across the training, internal validation, and external validation sets (Supplementary Figure 1, Figure 3). The ExtraTrees algorithm demonstrated the most balanced and stable performance across all three sets, achieving fusion AUCs of 0.865, 0.839, and 0.785, respectively, and was therefore selected as the final modeling algorithm.Using the optimal ExtraTrees algorithm, three models were constructed: a clinical-only model, an ITH-only model, and a combined clinical-ITH model. The ROC curves for these three models across the three cohorts are shown in Figure 4A. The combined model significantly outperformed both single-modality models in the training and internal validation sets. In the external validation set, the combined model (AUC = 0.785) achieved comparable performance to the clinical model (AUC = 0.790), while substantially outperforming the ITH-only model (AUC = 0.548). Calibration curves (Figure 4B) demonstrated good calibration for all three ExtraTrees models across the training, internal, and external cohorts, with predicted probabilities closely matching the observed MSI rates. Decision curve analysis (Figure 4C) indicated that, across clinically relevant threshold probability ranges, the combined model consistently yielded higher net benefit than either the clinical-only or ITH-only models, suggesting superior clinical utility.

2.4 SHAP Interpretability Analysis

To elucidate the decision-making mechanism of the combined ExtraTrees model, SHAP analysis was performed on the external validation set. The SHAP feature importance ranking (Figure 5A) revealed that both clinical-pathological and ITH-derived features made substantial contributions to model predictions. Tumor location was the highest-contributing feature, followed by T stage. Among radiomic features, ITH signature showed the greatest contribution,

followed sequentially by ITH product, ITH mean, ITH minimum, and ITH absolute difference. The individual ITH scores from single sequences (DWI ITH score and T2WI ITH score) contributed less than the multi-dimensional composite features.When SHAP values were aggregated by feature origin (Figure 5B), tumor location, T stage, and ITH signature accounted for 37.1%, 33.5%, and 29.3% of the total contribution, respectively, collectively representing the principal determinants of model decisions. Notably, ITH signature was the sole imaging-derived feature included in the grouped contribution analysis, and its contribution was comparable to that of T stage and only marginally lower than tumor location. SHAP force plots illustrating the individual prediction explanations for one MSI and one MSS patient are presented in Figure 6A and 6B

2.5 Bioinformatics Analysis

A combined transcriptomic dataset of TCGA-COAD and TCGA-READ was used for analysis. After quality control of clinical annotations and filtering for MSI status, a total of 271 colorectal cancer samples were included, comprising 82 in the MSI group and 189 in the MSS group (199 COAD and 72 READ samples). Following the filtration of lowly expressed genes, differential expression analysis was performed using DESeq2 with cancer type (COAD/READ) included as a covariate. A total of 378 significantly differentially expressed genes (DEGs) were identified using thresholds of $|\log_2 \text{fold change}| > 1$ and adjusted P-value < 0.05, of which 207 were upregulated and 171 were downregulated in the MSI group relative to the MSS group.
After variance-stabilizing transformation (VST) and row-wise Z-score normalization of the expression matrix, a clustering heatmap was generated (Figure 7A), demonstrating that MSI and MSS samples could be largely separated based on their differential gene expression profiles, suggesting distinguishable transcriptomic molecular features between the two tumor subtypes. A volcano plot (Figure 7B) illustrated the overall distribution of DEGs, with multiple upregulated genes reaching statistical significance.The protein sequences corresponding to all DEGs were imported into the STRING database (version 11.5) to construct a protein-protein interaction (PPI) network using a high-confidence interaction threshold (combined score ≥ 0.7). The PPI network data were imported into Cytoscape software (version 3.10.3) for visualization and topological analysis. The Maximal Clique Centrality (MCC) algorithm in the cytoHubba

plugin was applied to rank hub genes based on network centrality. The results (Figure 8) identified AKT1 as the core hub gene with the highest degree of connectivity within the PPI network.

2.6 Experimental Validation

Ten rectal cancer FFPE tissue specimens were included for experimental validation, comprising 5 MSS and 5 MSI-H cases. AKT1 protein expression in the two groups was assessed by immunohistochemistry (Figure 9) and quantified using H-score. Descriptive statistics showed that the mean H-score of AKT1 in the MSS group was 143.60 ± 26.14 (median: 132.02), compared with 76.74 ± 38.43 (median: 88.38) in the MSI-H group. The Shapiro-Wilk test confirmed normality of data distribution in both groups (MSS: P = 0.550; MSI-H: P = 0.125), and Levene's test indicated homogeneity of variances (F = 0.123, P = 0.735). The independent-samples t-test demonstrated that AKT1 protein expression was significantly higher in the MSS group than in the MSI-H group (t = 3.127, P = 0.012), suggesting a biologically meaningful difference in AKT1 protein expression between the two groups.

3 Discussion

Reliable methods for the non-invasive prediction of MSI status in rectal cancer based on MRI-derived intratumoral heterogeneity quantification remain lacking. In this study, we introduced an ITH scoring approach to quantitatively characterize spatial heterogeneity within rectal tumors using multicenter preoperative MRI data, and integrated ITH features with clinicopathological variables to develop a machine learning-based predictive model. The combined clinical-ITH model significantly outperformed both the clinical-only and ITH-only models in the training and internal validation sets (AUC = 0.865 and 0.839, respectively), while maintaining robust performance in the external validation set (AUC = 0.785). SHAP interpretability analysis further elucidated the differentiated contributions of key ITH features to MSI prediction. Building upon these imaging findings, we further integrated TCGA transcriptomic analysis with local immunohistochemical validation to explore molecular differences between MSI-H and MSS rectal cancers at the transcriptional and protein levels, providing exploratory clues for the molecular underpinnings of imaging phenotypic heterogeneity.

A close biological association exists between MSI status and intratumoral heterogeneity. MSI typically arises from defects in DNA mismatch repair (MMR) genes, leading to genomic instability with abundant insertions and deletions that drive pronounced tumor cell diversity [20,21]. Concurrently, MSI-H tumors exhibit distinctive tumor microenvironment features, including a high density of tumor-infiltrating lymphocytes, characteristic histopathological patterns (e.g., focal crypt distortion, Crohn-like lymphoid reaction), and significant spatial heterogeneity within the immune microenvironment [22,23]. These features result in the formation of functionally distinct subregions within MSI-H tumors that differ from those in MSS tumors: some areas are dominated by active immune infiltration, while others are driven primarily by tumor proliferation. Such intratumoral spatial heterogeneity provides a biological rationale for the imaging-based quantification of heterogeneity.

Previous CT- or MRI-based radiomics models have achieved notable progress in predicting MSI status in colorectal cancer. In the CT domain, Pei et al. [24]constructed a nomogram combining quantitative texture features with clinical factors based on preoperative multi-center CT (validation set AUC = 0.77), while Cao et al. [25] further incorporated radiomic features from the delayed phase of triphasic contrast-enhanced CT, substantially improving predictive performance (AUC = 0.964). In the MRI domain, Li et al. [26] developed a random forest model based on multi-parametric MRI, with external validation confirming the superior generalizability of the contrast-enhanced T1WI model (AUC = 0.77); Zhang et al. [27] compared five machine learning algorithms for MRI-based radiomics modeling, among which the logistic regression model achieved AUCs of 0.827 and 0.739 in the training and test sets, respectively. However, these studies share two common limitations: first, most employ whole-tumor mean features as the analytical unit, compressing spatial gradient information and failing to capture the regional distribution of MSI-related pathological changes; second, they largely rely on "black-box" model outputs without in-depth exploration of the biological mechanisms underlying key imaging features. Although Ma et al.[28] were the first to approach this question from a spatial regional perspective by constructing multiple CT models using a margin-inset strategy to differentiate tumor core from peripheral regions (AUC = 0.928), preliminarily demonstrating differential contributions of spatial feature distributions to MSI prediction, their geometric partitioning strategy constituted a predefined hard segmentation and did not provide

feature-level interpretability analysis. In contrast, the present study constructed a predictive model based on a multicenter cohort and introduced SHAP interpretability analysis to systematically quantify the differentiated contributions of individual features to MSI prediction, addressing the insufficient model transparency of prior studies.

Subregional radiomics (also known as habitat analysis) offers a methodological alternative to overcome the limitations of conventional whole-tumor analysis [29,30]. Its core advantage lies in the deconstruction of tumors into functional subregions with distinct imaging phenotypes through unsupervised clustering, thereby preserving spatial gradient information within the tumor rather than compressing it into a single scalar value. Wang et al. [31] constructed a habitat model using K-means clustering based on T2WI and DWI sequences, which not only outperformed whole-tumor analysis (training set AUC = 0.908, independent test set AUC = 0.863) but also revealed the differential contributions of different subregions to MSI prediction through SHAP analysis, establishing an initial link between spatial heterogeneity quantification and molecular subtyping. Cai et al. [32] further validated this paradigm in a multicenter cohort using a two-step clustering strategy combining fuzzy C-means and Bayesian Gaussian mixture models, achieving non-invasive localization of MSI-high-risk subregions and providing potential targets for imaging-guided biopsy. However, conventional K-means clustering imposes hard, mutually exclusive partitions on tumors, potentially losing critical heterogeneity information in transitional zones between subregions. Furthermore, independent analysis of each subregion followed by mechanical aggregation fails to capture overall tumor heterogeneity from a global perspective. The present study extends upon these methods: rather than relying on hard segmentation, the ITH score quantifies the spatial dispersion of feature values across the entire tumor from a global perspective, achieving continuous quantification of intratumoral heterogeneity. This approach avoids the information loss inherent in hard clustering while simultaneously accounting for both the global and regional aspects of tumor heterogeneity. This ITH scoring framework has demonstrated superior predictive performance over conventional radiomics in breast cancer and lung cancer [18,19,33], and the present study extends its application to the field of MSI prediction in rectal cancer.

Regarding model performance, this study systematically compared the predictive efficacy of clinical, ITH, and combined models across six machine learning algorithms. In the training

set, the combined model achieved an AUC of 0.865, significantly outperforming the clinical-only model (AUC = 0.757) and the ITH-only model (AUC = 0.776); in the internal validation set, the combined model achieved an AUC of 0.839, similarly surpassing the clinical model (AUC = 0.756) and the ITH model (AUC = 0.749). These findings indicate a complementary effect between ITH and clinical features — ITH features effectively captured MSI-associated imaging heterogeneity information, while clinical features provided robust cross-center generalizability. In the external validation set, the combined model (AUC = 0.785) achieved comparable performance to the clinical model (AUC = 0.790) while substantially outperforming the ITH model (AUC = 0.548), suggesting that when cross-center distribution shifts occur, the stability advantage of clinical baseline features becomes more prominent. The fusion of both feature types maximized training-phase performance while maintaining generalizability. Multivariate logistic regression identified age, tumor length, tumor location, and N stage as independent predictors of MSI status (all $P < 0.05$), consistent with previous reports[34,35]. MSI-type rectal cancer exhibits a distinctive clinical phenotype characterized by younger patient age, larger tumor volume, predilection for the upper rectum, and lower risk of lymph node metastasis [36-39]. Notably, tumor location was the highest-contributing feature in the SHAP analysis (37.1%), corroborating its identification as an independent predictor in the multivariate analysis and further confirming its importance in MSI prediction. The ITH score accounted for 29.3% of the SHAP contribution, demonstrating that MRI-quantified intratumoral heterogeneity carries independent predictive information beyond clinical factors, rather than merely serving as an appendage to clinical variables. SHAP dependence plots further revealed nonlinear interaction effects between features and prediction outcomes, providing interpretable quantitative evidence for understanding how these factors synergistically influence MSI status. From a clinical translation perspective, SHAP analysis transforms model predictions from a "black box" into a transparent process, enabling clinicians to trace the primary drivers of each individual prediction and thereby enhancing trust and confidence in model outputs.

Regarding mechanistic exploration, this study further investigated the differential mechanisms between MSI-H and MSS rectal cancers at the molecular biology level, aiming to provide a molecular basis for imaging phenotypic heterogeneity. Through the integration of

public database analysis and local sample validation, the following findings were obtained. First, differential expression analysis based on the TCGA database identified 378 DEGs between MSI and MSS tumors. Through PPI network topological analysis with AKT1 as the central hub gene, five representative DEGs (CAMK4, ITPKB, GPI, CKMT1A, and PCSK2) that directly interacted with AKT1 and showed the most significant expression differences were selected for subsequent discussion. Subsequently, immunohistochemical validation using five paired MSI-H and MSS rectal cancer specimens confirmed that AKT1 protein expression was significantly higher in MSS tumors than in MSI-H tumors ($P < 0.01$).

AKT1 is a core functional molecule of the PI3K-AKT-mTOR signaling pathway, and multiple studies have confirmed the critical regulatory role of this pathway in colorectal cancer initiation and progression[40-42]. Sustained pathway activation promotes aberrant tumor cell proliferation, inhibits apoptosis, induces metabolic reprogramming, and enhances tumor invasion and metastasis [41,43-45]. Combined with our experimental findings, it may be postulated that the lower AKT1 expression and weaker PI3K-AKT-mTOR pathway activation in MSI-H tumors could represent a potential mechanism underlying the longer overall survival and lower risks of tumor recurrence and distant metastasis observed in this subtype, which is consistent with our clinical factor selection results: compared with MSS tumors, MSI-H rectal cancer was more likely to present as localized disease (N0, without lymph node metastasis) with a lower risk of regional lymph node involvement. In contrast, MSS colorectal cancer frequently exhibits PI3K-AKT-mTOR pathway overactivation, with patients more commonly diagnosed at advanced disease stages and poorer overall prognosis. Furthermore, in MSS tumors, PI3K-AKT-mTOR pathway overactivation may maintain the integrity of the MMR-mTOR apoptotic pathway required for 5-FU-mediated cytotoxicity, conferring 5-FU chemosensitivity. Conversely, in MSI-H/dMMR tumors, insufficient pathway activation leads to inactivation of the MMR-mTOR apoptotic pathway and 5-FU chemoresistance; however, the mismatch repair deficiency results in accumulation of high mutational burden and abundant generation of tumor neoantigens, rendering these tumors more sensitive to immune checkpoint inhibitor therapy.

It is worth noting that the present study conducted an exploratory association analysis between imaging-based prediction and molecular mechanisms. Based on the existing literature

and our findings, the following hypothesis may be proposed: in MSI-H tumors, the relatively lower PI3K-AKT-mTOR pathway activation may attenuate tumor proliferation-driven growth, while immune infiltration-mediated regional histological changes (e.g., lymphoid reactions, mucinous degeneration) become more prominent, potentially leading to increased imaging signal differences across intratumoral regions and higher ITH scores. Conversely, in MSS tumors, PI3K-AKT-mTOR pathway overactivation may drive tumor cell proliferation as the dominant process, potentially resulting in more homogeneous intratumoral signals and lower ITH scores. However, the above imaging phenotype–molecular mechanism association remains hypothesis-driven reasoning based on the literature, lacking direct data-driven evidence (such as quantitative correlation analysis between ITH scores and AKT1 protein expression levels). The establishment of direct quantitative associations between imaging features and molecular markers awaits validation through larger sample sizes and multi-omics integrative analyses, representing a key direction for future investigation.

This study has several limitations. First, as a retrospective multicenter study, the prevalence of MSI positivity may differ between the training and validation sets due to population selection bias; although SMOTE was applied to balance the training data, synthetic samples cannot fully reproduce the full spectrum of imaging heterogeneity of true MSI tumors. Second, only T2WI and DWI sequences were included without integration of contrast-enhanced T1WI or other multi-parametric information, and multi-omics data such as genomics or pathomics were not incorporated, resulting in incomplete assessment of tumor angiogenesis and microstructural complexity. Third, manual tumor segmentation is time-consuming and subject to inter-observer variability; future studies may incorporate deep learning-based automatic segmentation algorithms to improve efficiency and reproducibility. Fourth, the sample sizes for bioinformatics analysis and experimental validation were relatively limited. The IHC validation included only 5 cases per group; although the between-group difference reached statistical significance ($P = 0.012$), the statistical power under such a small sample is insufficient, and 1–2 cases with altered values could potentially reverse the direction of results. Furthermore, the small sample size precludes reliable control for potential confounding factors such as age, sex, and tumor stage. The robustness of these findings requires validation in larger sample sizes. Fifth, imaging features (ITH scores) and molecular markers (e.g., AKT1 protein

expression) have not been directly quantitatively correlated; the relationship between imaging phenotypic heterogeneity and molecular pathway activation currently remains hypothesis-driven inference based on the literature, lacking data-driven statistical evidence, which needs to be addressed in future studies through correlation analyses. Sixth, selection bias inherent to retrospective design is unavoidable, and the clinical utility of the model requires further validation in prospective cohort studies.

4 Conclusion

In summary, this study developed and validated a rectal cancer MSI prediction model based on intratumoral heterogeneity scoring using multicenter preoperative MRI data. The quantification of intratumoral spatial heterogeneity through ITH scoring significantly improved predictive performance, and SHAP interpretability analysis revealed the independent contribution of ITH features to MSI prediction. Furthermore, through the integration of TCGA transcriptomic analysis and local sample validation, this study preliminarily revealed differential AKT1 expression between MSI-H and MSS rectal cancers within the PI3K-AKT-mTOR signaling pathway, providing exploratory clues for the molecular interpretation of imaging phenotypic heterogeneity. However, direct quantitative associations between imaging features and molecular markers have not yet been established, and the mechanistic findings are limited by the small sample size. Future studies with larger sample sizes, supplemented by imaging–molecular correlation analyses and prospective cohort designs, are warranted to further validate the generalizability and mechanistic hypotheses of the model.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.https://doi.org/10.3322/caac.21492
2. Xia C, Dong X, Li H, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (Engl) 2022;135(5):584–590.https://doi.org/10.1097/cm9.0000000000002108
3. De' Angelis GL, Bottarelli L, Azzoni C, et al. Microsatellite instability in colorectal

cancer. Acta Biomed 2018;89(9-s):97–101.https://doi.org/10.23750/abm.v89i9-S.7960
4. Battaglin F, Naseem M, Lenz HJ, Salem ME. Microsatellite instability in colorectal cancer: overview of its clinical significance and novel perspectives. Clin Adv Hematol Oncol 2018;16(11):735–745
5. André T, Shiu KK, Kim TW, et al. Pembrolizumab versus chemotherapy in microsatellite instability-high or mismatch repair-deficient metastatic colorectal cancer: 5-year follow-up from the randomized phase III KEYNOTE-177 study. Ann Oncol 2025;36(3):277–284.https://doi.org/10.1016/j.annonc.2024.11.012
6. Ribic CM, Sargent DJ, Moore MJ, et al. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N Engl J Med 2003;349(3):247–257.https://doi.org/10.1056/NEJMoa022289
7. Sargent DJ, Marsoni S, Monges G, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J Clin Oncol 2010;28(20):3219–3226.https://doi.org/10.1200/jco.2009.27.1825
8. Huang Q, Yu T, Li L, et al. Intraindividual Tumor Heterogeneity of Mismatch Repair Status in Metastatic Colorectal Cancer. Appl Immunohistochem Mol Morphol 2023;31(2):84–93.https://doi.org/10.1097/pai.0000000000001089
9. Amodio V, Lamba S, Chilà R, et al. Genetic and pharmacological modulation of DNA mismatch repair heterogeneous tumors promotes immune surveillance. Cancer Cell 2023;41(1):196–209.e195.https://doi.org/10.1016/j.ccell.2022.12.003
10. Sagaert X, Vanstapel A, Verbeek S. Tumor Heterogeneity in Colorectal Cancer: What Do We Know So Far? Pathobiology 2018;85(1-2):72–84.https://doi.org/10.1159/000486721
11. Saoudi González N, Salvà F, Ros J, et al. Unravelling the Complexity of Colorectal Cancer: Heterogeneity, Clonal Evolution, and Clinical Implications. Cancers (Basel) 2023;15(16).https://doi.org/10.3390/cancers15164020
12. Li S, Dai Y, Chen J, Yan F, Yang Y. MRI-based habitat imaging in cancer treatment: current technology, applications, and challenges. Cancer Imaging 2024;24(1):107.https://doi.org/10.1186/s40644-024-00758-9
13. Wu LX, Ding N, Ji YD, et al. Habitat Analysis in Tumor Imaging: Advancing Precision Medicine Through Radiomic Subregion Segmentation. Cancer Manag Res 2025;17:731–741.https://doi.org/10.2147/cmar.S511796
14. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77(21):e104–e107.https://doi.org/10.1158/0008-5472.Can-17-0339
15. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48(4):441–446.https://doi.org/10.1016/j.ejca.2011.11.036
16. He W, Huang W, Zhang L, Wu X, Zhang S, Zhang B. Radiogenomics: bridging the gap between imaging and genomics for precision oncology. MedComm (2020) 2024;5(9):e722.https://doi.org/10.1002/mco2.722
17. Zhang Y, Zhang X, Zhong X, et al. Immunophenotype-guided interpretable radiomics model for predicting neoadjuvant anti-PD-1 response in stage III-IV d-MMR/MSI-H

colorectal cancer. J Immunother Cancer 2025;13(8).https://doi.org/10.1136/jitc-2025-011569
18. Li J, Qiu Z, Zhang C, et al. ITHscore: comprehensive quantification of intra-tumor heterogeneity in NSCLC by multi-scale radiomic features. Eur Radiol 2023;33(2):893–903.https://doi.org/10.1007/s00330-022-09055-0
19. Huang Y, Wang X, Cao Y, et al. Nomogram for Predicting Neoadjuvant Chemotherapy Response in Breast Cancer Using MRI-based Intratumoral Heterogeneity Quantification. Radiology 2025;315(1):e241805.https://doi.org/10.1148/radiol.241805
20. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med 2015;372(26):2509–2520.https://doi.org/10.1056/NEJMoa1500596
21. Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. Gastroenterology 2008;135(4):1079–1099.https://doi.org/10.1053/j.gastro.2008.07.076
22. Jenkins MA, Hayashi S, O'Shea AM, et al. Pathology features in Bethesda guidelines predict colorectal cancer microsatellite instability: a population-based study. Gastroenterology 2007;133(1):48–56.https://doi.org/10.1053/j.gastro.2007.04.044
23. Rozek LS, Schmit SL, Greenson JK, et al. Tumor-Infiltrating Lymphocytes, Crohn's-Like Lymphoid Reaction, and Survival From Colorectal Cancer. J Natl Cancer Inst 2016;108(8).https://doi.org/10.1093/jnci/djw027
24. Pei Q, Yi X, Chen C, et al. Pre-treatment CT-based radiomics nomogram for predicting microsatellite instability status in colorectal cancer. Eur Radiol 2022;32(1):714–724.https://doi.org/10.1007/s00330-021-08167-3
25. Cao Y, Zhang G, Zhang J, et al. Predicting Microsatellite Instability Status in Colorectal Cancer Based on Triphasic Enhanced Computed Tomography Radiomics Signatures: A Multicenter Study. Front Oncol 2021;11:687771.https://doi.org/10.3389/fonc.2021.687771
26. Li Z, Zhang J, Zhong Q, et al. Development and external validation of a multiparametric MRI-based radiomics model for preoperative prediction of microsatellite instability status in rectal cancer: a retrospective multicenter study. Eur Radiol 2023;33(3):1835–1843.https://doi.org/10.1007/s00330-022-09160-0
27. Zhang Y, Liu J, Wu C, Peng J, Wei Y, Cui S. Preoperative Prediction of Microsatellite Instability in Rectal Cancer Using Five Machine Learning Algorithms Based on Multiparametric MRI Radiomics. Diagnostics (Basel) 2023;13(2).https://doi.org/10.3390/diagnostics13020269
28. Ma Y, Lin C, Liu S, 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 2022;12:1020349.https://doi.org/10.3389/fonc.2022.1020349
29. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. Radiology 2013;269(1):8–15.https://doi.org/10.1148/radiol.13122697
30. Waqar M, Van Houdt PJ, Hessen E, et al. Visualising spatial heterogeneity in glioblastoma using imaging habitats. Front Oncol

2022;12:1037896.https://doi.org/10.3389/fonc.2022.1037896
31. Wang Y, Xie B, Wang K, et al. Multi-parametric MRI Habitat Radiomics Based on Interpretable Machine Learning for Preoperative Assessment of Microsatellite Instability in Rectal Cancer. Acad Radiol 2025;32(7):3975–3988.https://doi.org/10.1016/j.acra.2025.02.009
32. Cai Z, Xu Z, Chen Y, et al. Multiparametric MRI subregion radiomics for preoperative assessment of high-risk subregions in microsatellite instability of rectal cancer patients: a multicenter study. Int J Surg 2024;110(7):4310–4319.https://doi.org/10.1097/js9.0000000000001335
33. Zhang J, Sha J, Liu W, Zhou Y, Liu H, Zuo Z. Quantification of Intratumoral Heterogeneity: Distinguishing Histological Subtypes in Clinical T1 Stage Lung Adenocarcinoma Presenting as Pure Ground-Glass Nodules on Computed Tomography. Acad Radiol 2024;31(10):4244–4255.https://doi.org/10.1016/j.acra.2024.04.008
34. 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.https://doi.org/10.1186/s12885-022-09584-3
35. Ma Y, Shi Z, Wei Y, Shi F, Qin G, Zhou Z. Exploring the value of multiple preprocessors and classifiers in constructing models for predicting microsatellite instability status in colorectal cancer. Sci Rep 2024;14(1):20305.https://doi.org/10.1038/s41598-024-71420-4
36. Gryfe R, Kim H, Hsieh ET, et al. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. N Engl J Med 2000;342(2):69–77.https://doi.org/10.1056/nejm200001133420201
37. Mármol I, Sánchez-de-Diego C, Pradilla Dieste A, Cerrada E, Rodriguez Yoldi MJ. Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. Int J Mol Sci 2017;18(1).https://doi.org/10.3390/ijms18010197
38. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med 2003;348(10):919–932.https://doi.org/10.1056/NEJMra012242
39. Gafà R, Maestri I, Matteuzzi M, et al. Sporadic colorectal adenocarcinomas with high-frequency microsatellite instability. Cancer 2000;89(10):2025–2037
40. Baba Y, Nosho K, Shima K, et al. Phosphorylated AKT expression is associated with PIK3CA mutation, low stage, and favorable outcome in 717 colorectal cancers. Cancer 2011;117(7):1399–1408.https://doi.org/10.1002/cncr.25630
41. Leiphrakpam PD, Are C. PI3K/Akt/mTOR Signaling Pathway as a Target for Colorectal Cancer Treatment. Int J Mol Sci 2024;25(6).https://doi.org/10.3390/ijms25063178
42. Johnson SM, Gulhati P, Rampy BA, et al. Novel expression patterns of PI3K/Akt/mTOR signaling pathway components in colorectal cancer. J Am Coll Surg 2010;210(5):767–776, 776–768.https://doi.org/10.1016/j.jamcollsurg.2009.12.008
43. Zhong J, Ding S, Zhang X, et al. To Investigate the Occurrence and Development of Colorectal Cancer Based on the PI3K/AKT/mTOR Signaling Pathway. Front Biosci (Landmark Ed) 2023;28(2):37.https://doi.org/10.31083/j.fbl2802037
44. Cohen MM, Jr. The AKT genes and their roles in various disorders. Am J Med Genet

A 2013;161a(12):2931–2937.https://doi.org/10.1002/ajmg.a.36101
45. Agarwal E, Brattain MG, Chowdhury S. Cell survival and metastasis regulation by Akt signaling in colorectal cancer. Cell Signal 2013;25(8):1711–1719.https://doi.org/10.1016/j.cellsig.2013.03.025

Table 1. Baseline clinicopathological characteristics of the study cohorts

Characteristic	Training cohort (n=198)	internal validation cohort (n=110)	External validation cohort (n=156)	P
Age (IQR)	60.00 (54.00,68.00)	62.50 (56.00,68.75)	62.50 (56.00,70.00)	0.08[a]
Tumor diameter(IQR)	51.00 (40.08,67.00)	50.45 (39.75.60.00)	50.00 (40.75,63.13)	0.34[a]
CA199(IQR)	11.30 (6.16,28.60)	11.35 (6.38,20.95)	15.68 (7.31,28.62)	0.26[a]
CA125(IQR)	9.21 (7.03,14.30)	10.30 (7.11,15.40)	14.74 (8.22,25.80)	0.38[a]
CEA(IQR)	3.21 (1.54,7.77)	3.41 (2.16,8.36)	4.85 (2.31,13.14)	0.07[a]
MSI status, n(%)				0.66[b]
MSS	172 (86.87)	98 (89.09)	133 (85.26)	
MSI	26 (13.13)	12(10.91)	23 (14.74)	
Gender, n(%)				0.31[b]
Male	116 (58.59)	63 (57.27)	102 (65.38)	
Female	82 (41.41)	47 (42.73)	54 (34.62)	
Tumor Location, n(%)				0.30[b]
Upper	63 (31.82)	35 (31.82)	35 (22.44)	
Middle	106 (53.54)	58 (52.73)	91 (58.33)	
Lower	29 (14.65)	17 (15.45)	30 (19.23)	
T stage, n(%)				0.04[b]
T1	2 (1.01)	3 (2.73)	1 (0.64)	
T2	20(10.10)	9 (8.18)	12 (7.69)	
T3	127 (64.14)	83 (75.45)	96 (61.54)	
T4	49 (24.75)	15 (13.64)	47 (30.13)	
N stage, n(%)				0.01[b]
N0	32 (16.16)	22(20.00)	47 (30.13)	
N1	66 (33.33)	29 (26.36)	35 (22.44)	
N2	100 (50.51)	59 (53.64)	74 (47.44)	
M stage, n(%)				0.28[b]
M0	192 (96.97)	106 (96.36)	146 (93.59)	
M1	6 (3.03)	4 (3.64)	10 (6.41)	

IQR: Interquartile range; MSI: Microsatellite instability; MSS: Microsatellite stability;

a: Kruskal-Wallis test； b: Chi-square test or Fisher’s exact test

Table 2. Univariable and multivariate logistic regression analysis of MSI Status in rectal cancer

Characteristic	Univariable		Multivariable	
	OR (95%CI)	P Value	OR (95%CI)	P Value
Gender				
Male	1.00 (Reference)			
Female	1.33 (0.67~2.64)	0.41		
Tumor location				
Upper	1.00 (Reference)		1.00 (Reference)	
Middle	0.53 (0.26 ~ 1.06)	0.07	0.69 (0.32~1.46)	0.33
Lower	0.08 (0.01~0.63)	**0.02**	0.11 (0.01 ~0.87)	**0.04**
T stage				
T1	1.00 (Reference)			
T2	0.51 (0.07 ~ 2.34)	0.325		
T3	0.62(0.12 ~ 3.28)	0.558		
T4	1867398.77 (0.00 ~Inf)	0.989		
N stage				
N0	1.00 (Reference)		1.00 (Reference)	
N1	0.17 (0.04~0.71)	**0.015**	0.21 (0.05~0.93)	**0.039**
N2	0.68 (0.25 ~ 1.84)	0.448	0.52 (0.17~1.59)	0.250
Age	0.94 (0.91~0.97)	**0.01**	0.94 (0.91 ~ 0.97)	**<0.01**
Tumor diameter	1.02 (1.01~1.03)	**0.02**	1.02 (1.01 ~ 1.04)	**0.05**
CA199	1.00(1.00~1.00)	0.99		
CA125	1.00 (0.97~1.02)	0.85		
CEA	1.00 (0.99 ~1.01)	0.88		

OR: Odds Ratio, CI: Confidence Interval. Bold values indicates statistical significant value of $p < 0.05$.

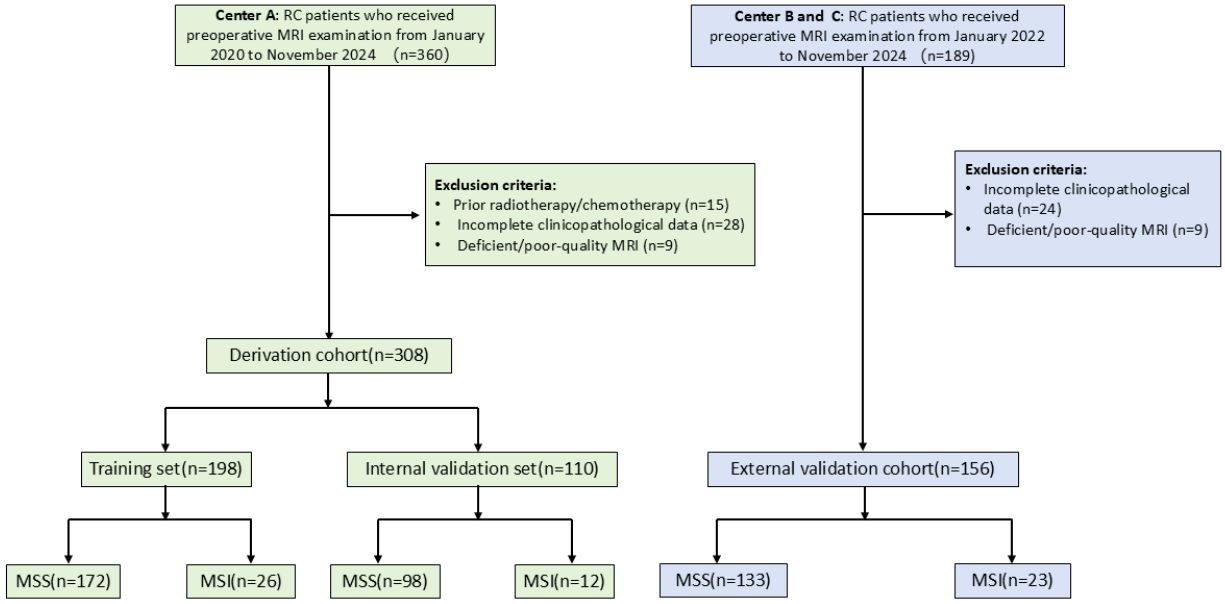


Figure1 Flowchart showing participant enrollment. RC: Rectal cancer

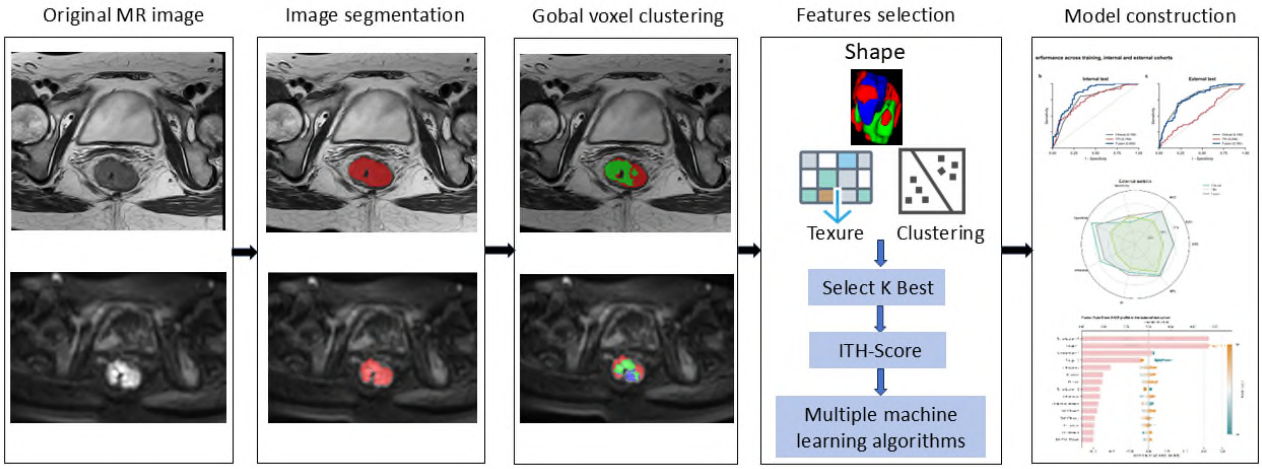


Figure 2. Schematic workflow of the present study. First, multi-parametric raw magnetic resonance (MR) images (T2WI and DWI sequences) of patients with rectal cancer were acquired. Second, manual segmentation of the whole-tumor volume of interest (VOI) was performed on MR images. Third, global voxel-wise K-means clustering was implemented on all tumor voxels to partition subregions with distinct intratumoral heterogeneity (ITH). Fourth, shape and texture radiomic features were extracted from each subregion, followed by feature screening to calculate the ITH-Score for each patient. Fifth, six machine learning algorithms were applied to establish three prediction models (clinical-only model, ITH-only model,

clinical-ITH fusion model). Model discrimination performance was validated in the training, internal validation and external validation cohorts, and SHapley Additive exPlanations (SHAP) analysis was conducted to interpret model prediction logic for comprehensive performance evaluation.

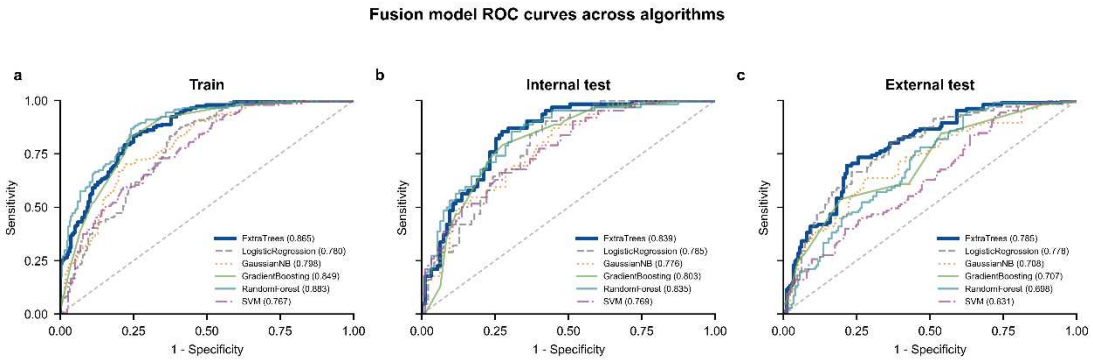


Figure 3 ROC curves of six machine learning algorithms based on the clinical-ITH fusion feature set in three cohorts.

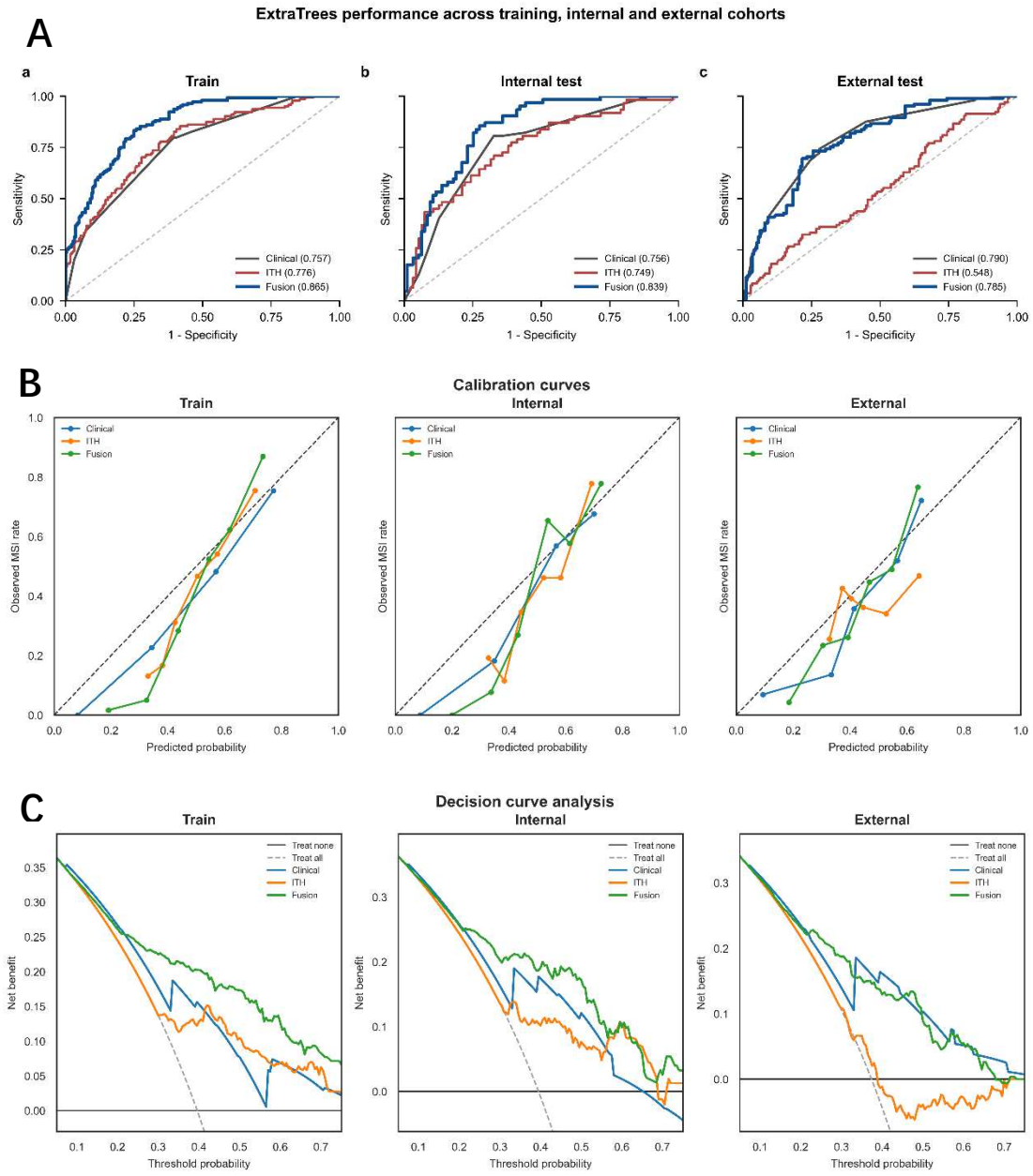


Figure 4 Comprehensive comparison of the three prediction models. (A) ROC curves of the clinical, ITH and fusion models constructed by the ExtraTrees algorithm in the training, internal validation and external validation cohorts. (B) Calibration curves of the three prediction models across the three cohorts. (C) Decision curve analysis of the three models in the training, internal validation and external validation cohorts.

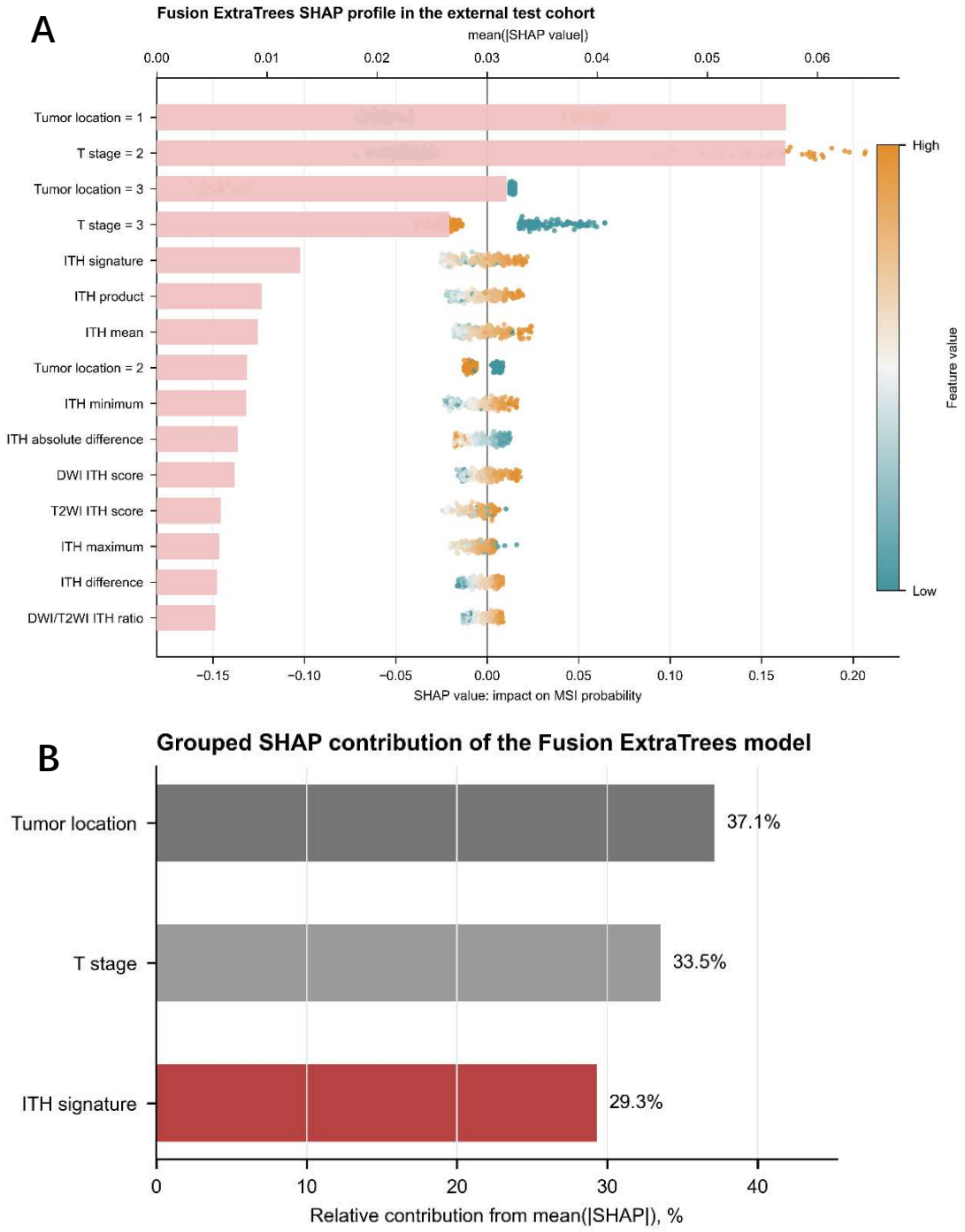


Figure 5 SHAP interpretability analysis of the fusion ExtraTrees model in the external validation cohort.(A) SHAP summary plot. The horizontal pink bar length represents the average absolute SHAP value, reflecting the global predictive contribution of each feature, sorted from top to bottom by descending contribution. The horizontal coordinate denotes the SHAP value, which indicates the positive or negative impact of each feature on the predicted

MSI probability: positive SHAP values increase the risk of MSI, while negative values reduce MSI risk. For categorical variables, Tumor location = 1/2/3 corresponds to upper/middle/lower rectal tumor; T stage = 2/3 corresponds to T2/T3 stage. The color gradient of scatter points corresponds to the original feature value, where orange represents high feature values and blue represents low feature values.(B) Bar chart of grouped relative SHAP contributions, quantifying the overall predictive weight of tumor location, T stage and ITH signature.

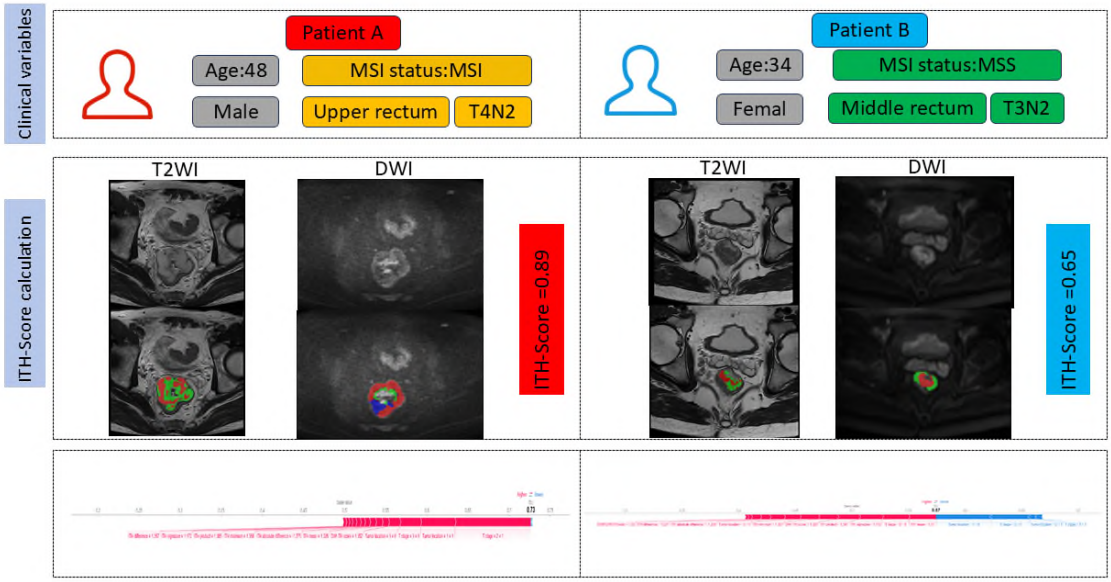


Figure 6 Two representative rectal cancer patients with different MSI status are presented, including clinical data, segmented T2WI and DWI MRI images for ITH-Score calculation, and individualized SHAP force plots. Patient A (MSI-H, 48-year-old male, upper rectum, T4N2) had a high ITH-Score of 0.89, and most ITH features positively elevated the MSI-H predictive probability. Patient B (MSS, 34-year-old female, middle rectum, T3N2) showed mild intratumoral heterogeneity with an ITH-Score of 0.65, and several ITH features reduced the risk of MSI-H prediction. Red bars indicate features increasing MSI-H probability, while blue bars represent features decreasing MSI-H probability.



Figure 7 Analysis of differentially expressed genes (DEGs) between MSI and MSS colorectal cancer samples. A: Clustering heatmap integrating TCGA-COAD (colon adenocarcinoma) and TCGA-READ (rectal adenocarcinoma) datasets. Rows are DEGs, columns are tumor samples; top bars annotate dataset source and MSI/MSS subtype. Colors show normalized gene expression (red: high expression; blue: low expression). B: Volcano plot comparing MSI versus MSS. x-axis: ($\log_2$(Fold Change), y-axis: -$\log_{10}$ (adjusted *P*-value). Red dots: significantly up-regulated genes in MSI group; blue dots: significantly down-regulated genes; gray dots: genes with no statistical significance.

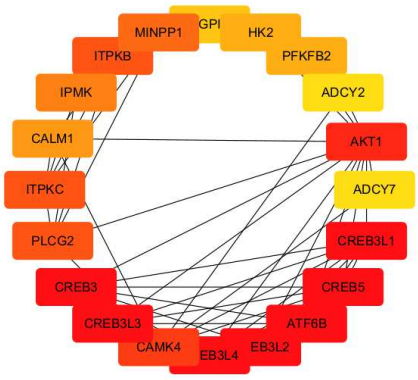
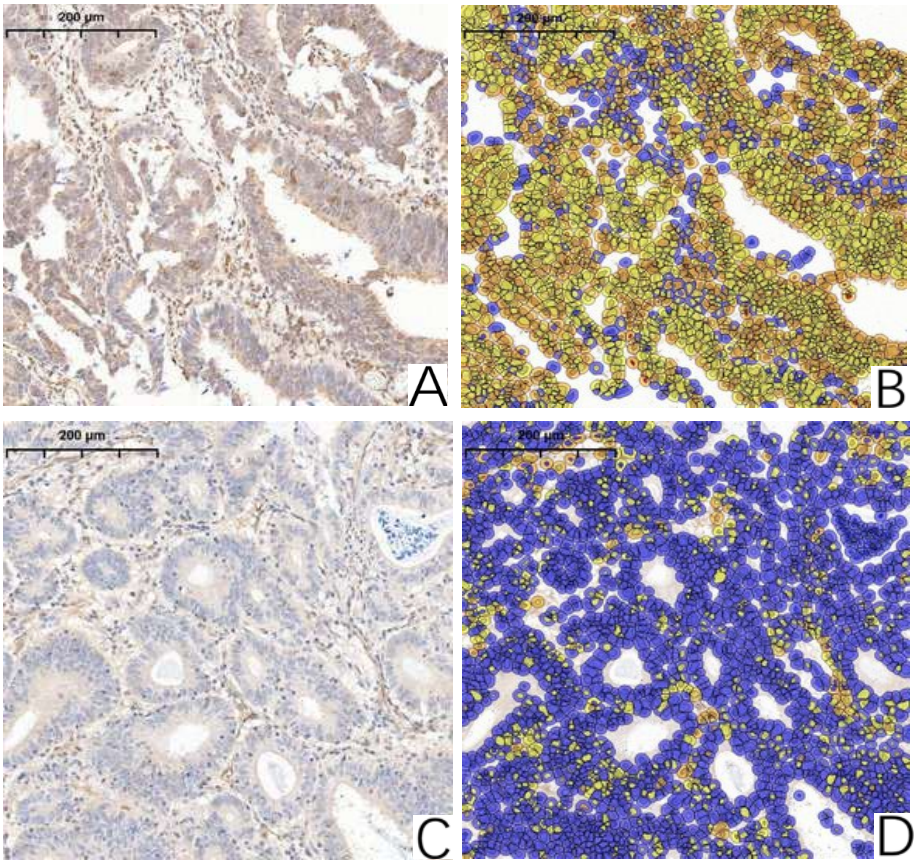


Figure 8 Protein-protein interaction (PPI) network of differentially expressed genes between MSI and MSS colorectal cancer. Nodes represent differentially expressed genes, and edges indicate protein-protein interactions. Nodes in red denote up-regulated genes in the MSI group, while yellow and orange nodes represent genes with different expression patterns.



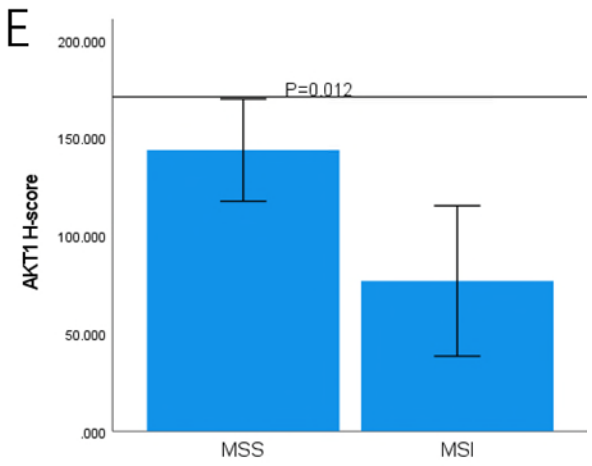


Figure 9 Immunohistochemical expression of AKT1 protein in MSI and MSS colorectal cancer tissues. A-B: Representative AKT1 immunohistochemical staining and corresponding analytical results of MSS-type colorectal cancer. C-D: Representative AKT1 immunohistochemical staining and corresponding analytical results of MSI-type colorectal cancer, scale bar = 200 μm. E: Comparison of AKT1 protein H-score between MSS and MSI groups. The H-score of AKT1 protein was significantly higher in the MSS group than in the MSI group, $P = 0.012$.

Supplementary Files

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

- SupplementaryFiles.docx