



Non-invasive prediction of microsatellite instability in colorectal cancer by a genetic algorithm–enhanced artificial neural network–based CT radiomics signature

Xiaobo Chen^{1,2,3} · Lan He^{1,2} · Qingshu Li⁴ · Liu Liu⁵ · Suyun Li^{1,2,6} · Yuan Zhang^{1,2,3} · Zaiyi Liu^{1,2} · Yanqi Huang^{1,2} · Yun Mao^{1,5} · Xin Chen⁷

Received: 25 February 2022 / Revised: 8 May 2022 / Accepted: 8 June 2022 / Published online: 30 June 2022
© The Author(s), under exclusive licence to European Society of Radiology 2022

Abstract

Objective The stratification of microsatellite instability (MSI) status assists clinicians in making treatment decisions for colorectal cancer (CRC) patients. This study aimed to establish a CT-based radiomics signature to predict MSI status in patients with CRC.

Methods A total of 837 CRC patients who underwent preoperative enhanced CT and had available MSI status data were recruited from two hospitals. Radiomics features were extracted from segmented tumours, and a series of data balancing and feature selection strategies were used to select MSI-related features. Finally, an MSI-related radiomics signature was constructed using a genetic algorithm–enhanced artificial neural network model. Combined and clinical models were constructed using multivariate logistic regression analyses by integrating the clinical factors with or without the signature. A Kaplan–Meier survival analysis was conducted to explore the prognostic information of the signature in patients with CRC.

Results Ten features were selected to construct a signature which showed robust performance in both the internal and external validation cohorts, with areas under the curves (AUC) of 0.788 and 0.775, respectively. The performance of the signature was comparable to that of the combined model (AUCs of 0.777 and 0.767, respectively) and it outperformed the clinical model constituting age and tumour location (AUCs of 0.768 and 0.623, respectively). Survival analysis demonstrated that the signature could stratify patients with stage II CRC according to prognosis (HR: 0.402, $p = 0.029$).

Conclusions This study built a robust radiomics signature for identifying the MSI status of CRC patients, which may assist individualised treatment decisions.

Xiaobo Chen, Lan He, and Qingshu Li contributed equally to this work.

✉ Yanqi Huang
yikiann@126.com

✉ Yun Mao
maoyun1979@163.com

✉ Xin Chen
wolfchenxin@163.com

¹ Department of Radiology, Guangdong Provincial People’s Hospital, Guangdong Academy of Medical Sciences, 106 Zhongshan Er Road, Guangzhou 510080, China

² Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People’s Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China

³ The Second School of Clinical Medicine, Southern Medical University, Guangzhou 510515, China

⁴ Department of Pathology, College of Basic Medicine, Chongqing Medical University, Chongqing 400016, China

⁵ Department of Radiology, the First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, China

⁶ School of Medicine, South China University of Technology, Guangzhou 510006, China

⁷ Department of Radiology, Guangzhou First People’s Hospital, School of Medicine, South China University of Technology, 1 Panfu Road, Guangzhou 510180, China

Key Points

- *Our well-designed modelling strategies helped overcome the problem of data imbalance caused by the low incidence of MSI.*
- *Genetic algorithm–enhanced artificial neural network–based CT radiomics signature can effectively distinguish the MSI status of CRC patients.*
- *Kaplan–Meier survival analysis demonstrated that our signature could significantly stratify stage II CRC patients into high- and low-risk groups.*

Keywords Colorectal neoplasms · Microsatellite instability · Neural networks · Survival analysis

Abbreviations

ANN	Artificial neural network
CEA	Carcinoembryonic antigen
CRC	Colorectal cancer
GANN	Genetic algorithm–enhanced artificial neural network
IHC	Immunohistochemistry
MMR	Mismatch repair
MSI	Microsatellite instability
MSI-H	High MSI
MSI-L	Low MSI
MSS	Microsatellite stability
PCA	Principal component analysis
PCR	Polymerase chain reaction
RFE	Recursive feature elimination
SVM	Support vector machine
VOI	Volume of interest

Introduction

Colorectal cancer (CRC) is driven by numerous genetic and epigenetic events, and it has the third-highest mortality rate worldwide [1]. Microsatellite instability (MSI) is a major oncogenic factor that is defined as the generalised instability of short tandem repeat DNA sequences (microsatellites), which occurs in approximately 15% of CRC cases. Many studies have confirmed that MSI is a promising biomarker for the diagnosis of Lynch syndrome diagnosis [2], for predicting the treatment response [3] and prognosis of CRC patients [4, 5].

The National Comprehensive Cancer Network guidelines recommend MSI testing for patients [6]. MSI can be detected using polymerase chain reaction (PCR) or immunohistochemistry (IHC) for mismatch repair (MMR) protein expression. However, an invasive biopsy may cause a potential risk of complications such as bleeding and bowel perfusion. More importantly, the degree of MMR protein expression detected at different biopsy locations may be underestimated or overestimated owing to tumour heterogeneity [7]. Therefore, it is necessary to seek a non-invasive method to provide additional diagnostic information pertaining to MSI status prior to the initiation of procedures and treatments [8].

Computed tomography (CT) is a common imaging method that plays an important role in CRC staging; however, it is considerably challenging for radiologists to assess MSI status based on macroscopic CT images. Radiomics, a method of extracting high-throughput qualitative features from routine medical radiologic images, has the potential to characterise tumour phenotypes and improve cancer diagnosis, prognosis, and response to therapy [9, 10]. Several studies have suggested that it is feasible to predict the MSI status of CRC using radiomics methods [11–18]. However, most previous studies had modest datasets (less than 200) and lacked independent external validation cohorts [14–18]. Additionally, data imbalance caused by the low incidence of MSI hinders accurate predictive analysis in real-world applications [19]. Although a 1:1 or 2:1 random matching method has been performed in several previous studies [13, 15–18], a possible selection bias might limit real-world applications [20]. Moreover, all previous studies focused only on assessing the value of radiomics for MSI status prediction and failed to provide further prognostic information.

In this study, we aimed to explore the potential diagnostic value of MSI status and prognostic stratification using a CT-based radiomics approach based on real-world data in CRC patients.

Materials and methods

Patients and data collection

This retrospective study was approved by the Institutional Review Board of Guangdong Provincial People’s Hospital and the First Affiliated Hospital of Chongqing Medical University. The requirement for informed consent forms was waived due to the retrospective nature of the study. Consecutive patients pathologically diagnosed with CRC were recruited from two participating hospitals between May 2005 and July 2019.

The inclusion and exclusion processes are illustrated in Fig. 1. Patients with CRC enrolled in hospital I were randomly stratified into two independent cohorts (70% of patients were enrolled in the training cohort and 30% in the internal validation cohort) according to MSI status, and patients in hospital II were included in the external validation cohort. Overall

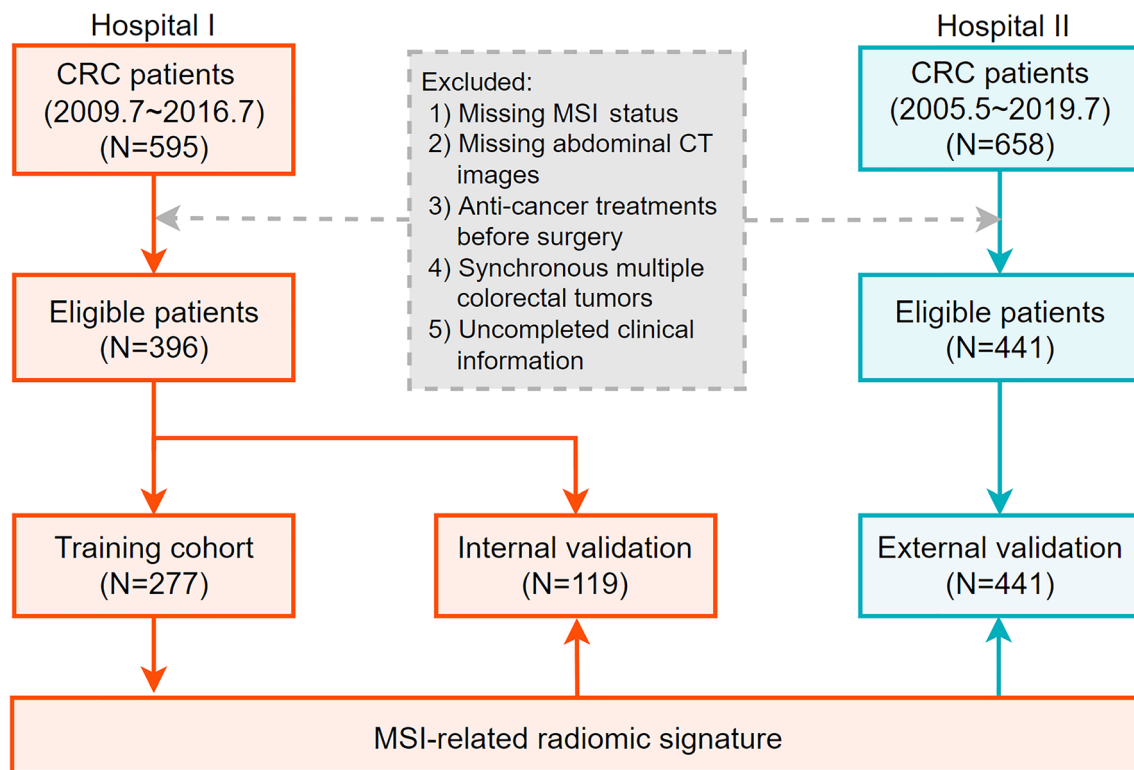


Fig. 1 Flowchart of inclusion and exclusion criteria for eligible patients in the study. CRC, colorectal cancer; MSI, microsatellite instability

survival analysis was performed on hospital I data, considering sufficient follow-up.

The patients' clinical data, including age, gender, tumour location, TNM stage, carcinoembryonic antigen (CEA), and Ki-67, were retrospectively reviewed and recorded from the medical record archives. Overall survival (OS) was defined as the time from surgery to death from any cause. Those who were alive at the time of the last follow-up were considered censored data.

Identification of MSI status

MSI status was assessed using IHC staining to detect the expression levels of MMR gene proteins (MLH1, MSH2, MSH6, and PMS2). The results were reviewed and confirmed by two pathologists with more than 10 years of experience in the diagnosis of abdominal disease. Patients were divided into MSI-L/microsatellite stability (MSS) and MSI-H groups: the former was positive for all four MMR proteins, while the latter included patients who were negative for any one of the MMR proteins.

CT image acquisition

CT images were obtained using six CT scanners from two institutions. All patients underwent enhanced abdominal CT scans. For hospital I, patients were examined using 8-slice (Lightspeed Ultra 8, GE Healthcare), 64-slice (LightSpeed VCT, GE Healthcare), or 256-slice (Brilliance iCT, Philips Healthcare) multidetector CT scanners. For hospital II, 64-

slice (Somatom Sensation, Siemens Healthcare or LightSpeed VCT, GE Healthcare) or 128-slice (Discovery CT750, GE Healthcare) CT scanners were used to perform abdominal CT scans. The acquisition parameters of the two institutions were as follows: tube voltage, 120 keV; tube current, 130–250 mAs, collimation, 0.625–2.5 mm, slice thickness, 1–5 mm, reconstruction interval, 0.625–1.25 mm. The contrast agents were injected at a speed of 2.5–3.5 mL/s with a high-pressure pump syringe. Portal phase CT scans were performed 55–75 s after the injection of the contrast agents.

Image pre-processing and tumour segmentation

To minimise the centre effect of CT images from different scanners and hospitals [21], all original CT images were appropriately pre-processed. First, the images were resampled to $1 \times 1 \times 1 \text{ mm}^3$ (x, y, z) using a linear interpolation algorithm to standardise the voxel spacing [22]. Then, a 25 HU bin width was set to discretise the voxel intensity and reduce the noise [23].

Tumours were segmented manually on the axial images at the portal venous phase by one radiologist with 10 years of experience with CRC diagnosis using ITK-SNAP software (version 2.2.0, www.itksnap.org), and mesentery air and pericolonic fat were excluded from the volume of interest (VOI) data. Another radiologist with 15 years of experience in abdominal disease evaluation revised and confirmed the segmentation results. Tumour VOIs were saved for subsequent quantitative feature extraction.

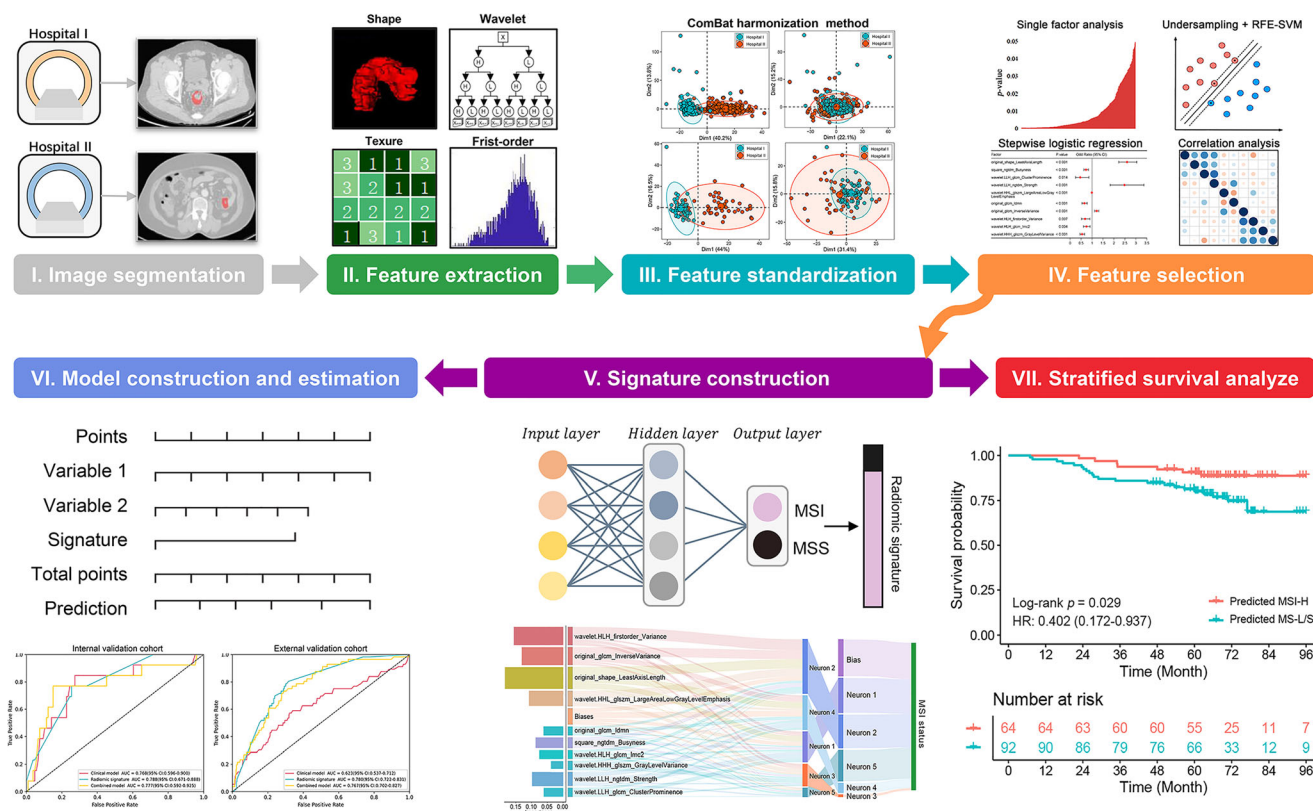


Fig. 2 Workflow for radiomics analysis in this study. (I) Tumour segmentation on CT images. (II) Radiomics features extraction from CT images. (III) For data pre-processing, the ComBat harmonisation method was used for feature standardisation. (IV) A series of coarse-to-fine feature selection strategies were carried out for feature selection. (V) The microsatellite instability (MSI)–related radiomics signature was

constructed with an enhanced neural network model. (VI) Building and performance of model: receiver operating characteristics (ROC) curves were implemented to assess the models' performance. Fitness curves and Sankey diagrams were plotted to visualise the network results. (VII) Prognostic analysis was performed based on the MSI-related radiomics signature

Radiomics feature extraction and standardisation

CT radiomics features of tumour VOI were automatically extracted using the PyRadiomics software package (version 3.0.1, <https://pyradiomics.readthedocs.io/en/latest/changes.html>), which included shape, first-order, texture, wavelet, exponential, and square transform features [24].

ComBaTool and z -scores were then used to normalise the features. ComBaTool, a free online application (https://forlhac.shinyapps.io/Shiny_ComBat/) [25], was used to pool features and minimise inter-scanner variability [26]. Principal component analysis (PCA) was utilised to visualise the effects of Combat on feature uniformisation. Finally, all radiomics features were standardised using z -scores. The radiomics analysis process is shown in Fig. 2.

Radiomics feature selection

As illustrated in Fig. 3, a series of coarse-to-fine data balancing and feature selection strategies were performed to deal with the imbalanced dataset and dimensionality reduction (Fig. 3A–D). Initially, an independent-sample t -test or Mann-Whitney U test

was performed to coarsely choose features with the statistical differences between the MSI-H and MSI-L/MSS groups. Then, the random under-sampling method generated K -balanced subsets with an equal number of MSI-H and MSI-L/MSS patients. A recursive feature elimination support vector machine (RFE-SVM) method was designed to recursively train these samples using a SVM model with the linear kernel function. Features with minimum weight were eliminated, and the top-ranking feature subset was retained. High-frequency features were then selected by counting the occurrences in the K features' subsets. Subsequently, upsampling was performed to alleviate model bias to majority class, followed by backward stepwise logistic regression to further reduced the features. Finally, a post hoc correlation analysis was used to inspect the redundancy of the final selected features.

Construction of the MSI-related radiomics signature

Artificial neural network (ANN) models are widely popular owing to their strong ability to describe complex and non-linear problems. Genetic algorithm-enhanced ANN (GANN) (Fig. 3E) has the potential to accelerate the updating of network

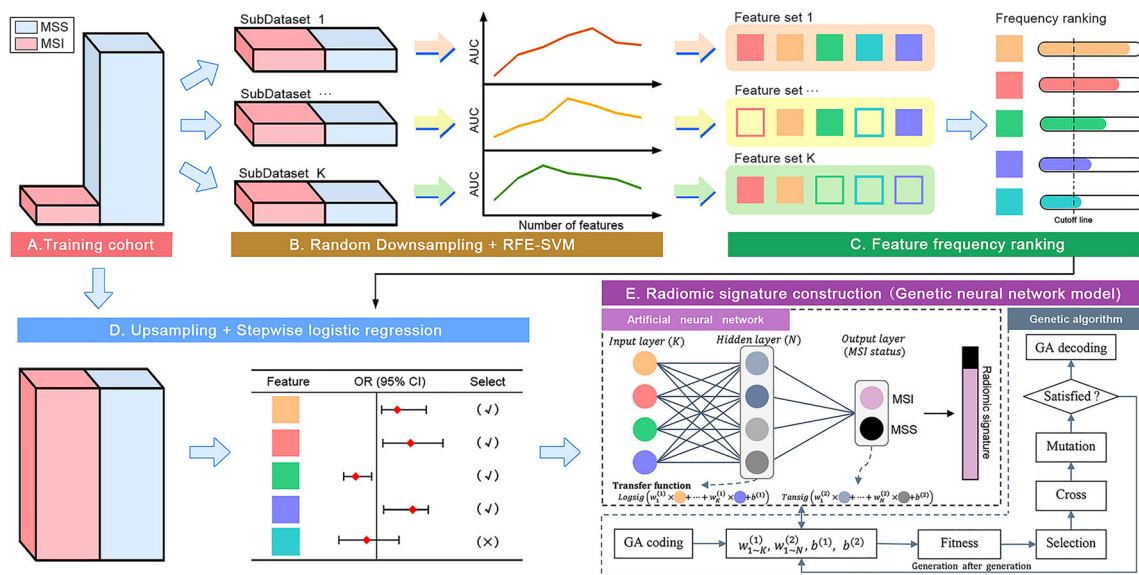
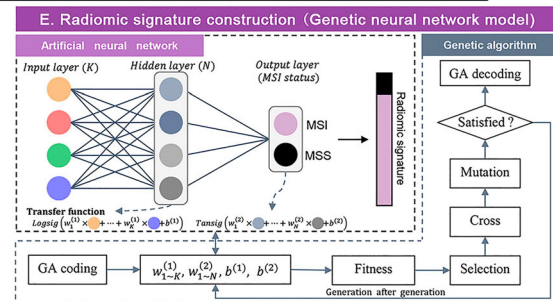


Fig. 3 Flowchart of feature selection and modelling strategies for imbalanced data handling. (A–C) High-frequency feature selection through a combination of random downsampling and recursive feature elimination support vector machine (RFE-SVM) ranking strategies. **D** Feature selection through upsampling and stepwise logistic regression. **E** The structure of the genetic algorithm–enhanced artificial neural



network. The upper left is the basic structure of the artificial neural network, with radiomics features as inputs and MSI statuses as outputs. Two activation functions are marked at the bottom. The genetic algorithm was responsible for decoding the network parameters and optimising them via the “natural selection” method. The output of the optimal model was regarded as the radiomics signature

parameters, avoid local optimisation, and improve prediction accuracy [27, 28].

The GANN model consisted of an input layer (radiomics features), a hidden layer, and an output layer (MSI status). Sigmoid and tangent activation functions were employed in the hidden and output layers, respectively. To prevent overfitting, the fitness function of the GA was described as $\min\{AUC_{train}, AUC_{inter-validation}\}$, and the best parameter result was decoded after repetitive generations of population selection, crossover, and mutation operators. The output of the optimal GANN model was converted into a radiomics signature.

Development and validation of the predictive models

Univariate and multivariate logistic regression analyses were performed to construct the predictive models. The candidate clinical factors were gender (male versus female), age (continuous), primary tumour location [left (splenic flexure to sigmoid) versus right (caecum to transverse colon) or rectum], CEA (continuous), and Ki-67 (continuous). Then, a combined model was built incorporating all significant predictive factors, and a clinical model was constructed using only significant clinical factors.

Statistical analysis

Chi-square, Fisher's exact, or Mann-Whitney *U* tests were performed on the clinical factors in the training, internal and external validation cohorts, and between the MSI-H and MSI-L/MSS groups. Receiver operator characteristic (ROC) curve and area

under the curve (AUC) values were used to evaluate the performance of the radiomics signature and prediction models. The DeLong test was used to compare AUC differences.

To evaluate the prognostic stratification value of the radiomics signature, patients were classified into predicted MSI-H and MSI-L/MSS groups by maximising their Youden index values. Survival curves were plotted using the Kaplan–Meier method and compared using the log-rank test. The GANN was operated in the MATLAB 2018b toolbox (MathWorks). Statistical analyses were performed using R (version 4.1.1) and Python (version 3.7.3). *p* values < 0.05 were considered statistically significant.

Results

Patient characteristics

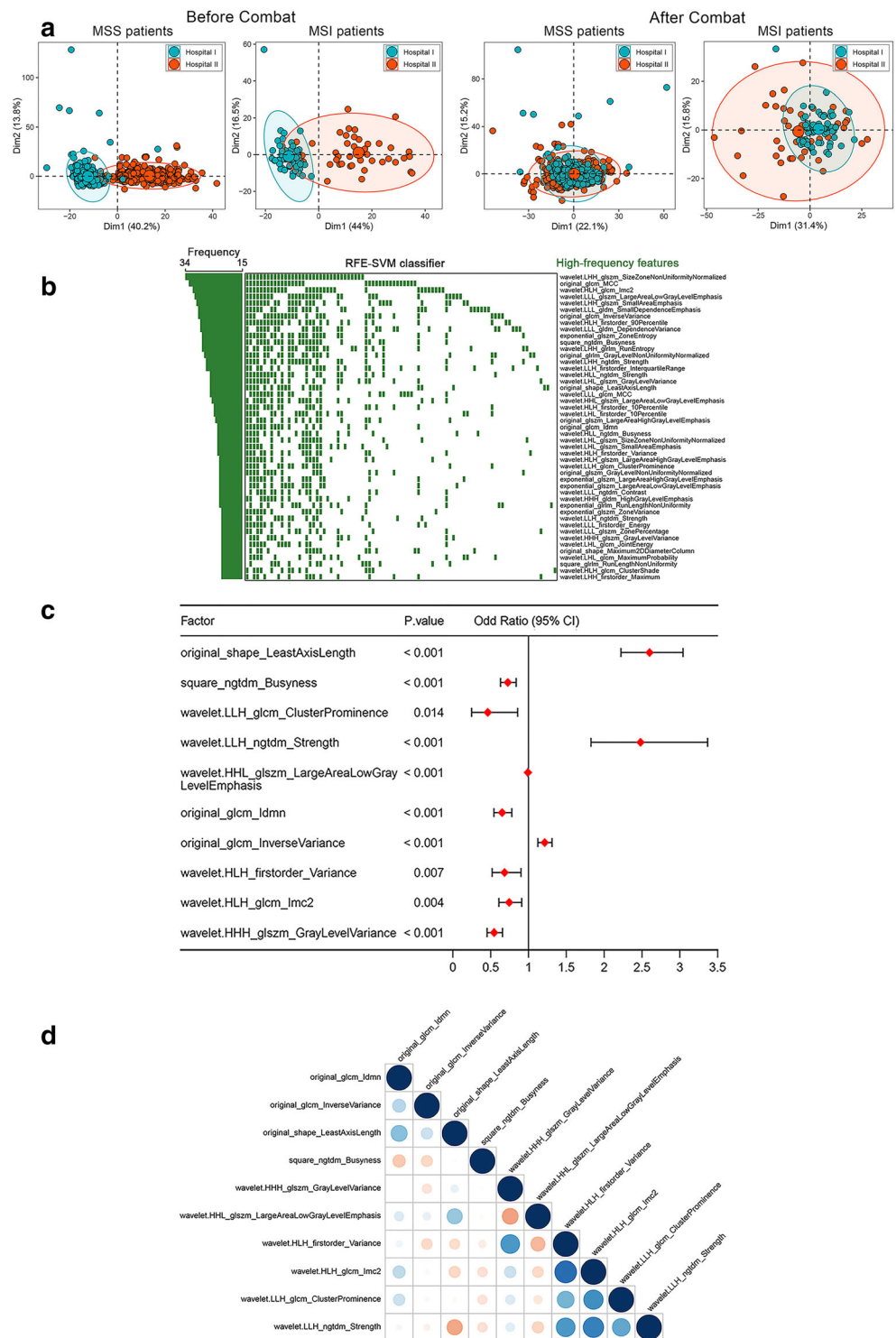
A total of 837 CRC patients with a MSI-H prevalence of 10.6% (42/396) in hospital I and 12.6% (56/441) in hospital II were included in the study. Table 1 shows the patients' clinical information and a comparison between patients with MSI-H and MSI-L/MSS. No statistical differences were found among all factors expect tumour location (*p* < 0.001). No variable was significantly different between the training and internal validation cohorts (*p* = 0.122~1.000). Moreover, 99.2% (393/396) of patients were followed up for OS, and the median [Q₂₅, Q₇₅] OS was 65.5 [50.3, 76.2] months for hospital I.

Table 1 Clinicopathologic characteristics of 837 CRC patients from two centres

Characteristics	Training cohort (N = 277)		Internal validation cohort (N = 119)		External validation cohort (N = 441)		p value
	MSI-L/MSS (N = 248)	MSI-H (N = 29)	MSI-L/MSS (N = 106)	MSI-H (N = 13)	MSI-L/MSS (N = 385)	MSI-H (N = 56)	
Demographics and clinical characteristics							
Age (median [IQR])	64.00 [55.00, 71.00]	56.00 [47.00, 71.00]	65.00 [58.25, 72.00]	60.00 [56.00, 74.00]	65.00 [56.00, 72.00]	62.50 [47.75, 72.25]	0.202
Gender (%)							0.432
Female	94 (37.9)	11 (37.9)	48 (45.3)	3 (23.1)	144 (37.4)	24 (42.9)	
Male	154 (62.1)	18 (62.1)	58 (54.7)	10 (76.9)	241 (62.6)	32 (57.1)	
Tumour location (%)							< 0.001
Left colon	69 (27.8)	10 (34.5)	32 (30.2)	1 (7.7)	97 (25.2)	10 (17.9)	
Rectum	123 (49.6)	4 (13.8)	50 (47.2)	2 (15.4)	202 (52.5)	20 (35.7)	
Right colon	56 (22.6)	15 (51.7)	24 (22.6)	10 (76.9)	86 (22.3)	26 (46.4)	
CEA (median [IQR])	3.00 [2.00, 8.00]	3.00 [2.00, 7.00]	4.00 [2.00, 12.00]	2.00 [2.00, 7.00]	4.00 [1.98, 8.00]	3.90 [1.80, 7.60]	0.897
Ki-67 (median [IQR])	0.80 [0.60, 0.90]	0.85 [0.60, 0.90]	0.80 [0.70, 0.90]	0.70 [0.60, 0.85]	0.60 [0.50, 0.70]	0.60 [0.48, 0.80]	0.356
Pathologic characteristics							
T stage (%)							0.056
T ₁	8 (3.2)	0 (0.0)	3 (2.8)	0 (0.0)	0 (0.0)	0 (0.0)	
T ₂	39 (15.7)	5 (17.2)	11 (10.4)	2 (15.4)	0 (0.0)	0 (0.0)	
T ₃	182 (73.4)	23 (79.3)	84 (79.2)	7 (53.8)	269 (69.9)	32 (57.1)	
T ₄	19 (7.7)	1 (3.4)	8 (7.5)	4 (30.8)	116 (30.1)	24 (42.9)	
N stage (%)							/
N ₀	137 (55.2)	20 (69.0)	51 (48.1)	6 (46.2)	385 (100.0)	56 (100.0)	
N ₁	73 (29.4)	4 (13.8)	38 (35.8)	3 (23.1)	0 (0.0)	0 (0.0)	
N ₂	38 (15.3)	5 (17.2)	17 (16.0)	4 (30.8)	0 (0.0)	0 (0.0)	
M stage (%)							/
M ₀	245 (98.8)	29 (100.0)	105 (99.1)	13 (100.0)	385 (100.0)	56 (100.0)	
M ₁	3 (1.2)	0 (0.0)	1 (0.9)	0 (0.0)	0 (0.0)	0 (0.0)	

Note. IQR is short for interquartile range. p value[†] represents the difference between training and internal validation cohort

Fig. 4 Radiomics feature selection. **a** Principal component analysis (PCA) scatter plot. The first two principal components of the significant radiomics features were visualised in a two-dimensional scatter plot before and after using ComBat. The *X* and *Y* axes present the two principal component dimensions. Red and blue colours represent the two hospitals. **b** Frequency distribution heatmap and histogram. One hundred feature subsets were generated by recursive feature elimination support vector machine (RFE-SVM) and drawn as a grid heatmap after frequency ranking. High-frequency features are marked on the right, and the corresponding frequency histogram is shown on the left. **c** Forest plot. Selected features and their *p* values are described in the table, and the odds ratios are plotted as line segments with red diamonds. **d** Correlation heatmaps. The blue circle represents a positive correlation, and the red indicates a negative correlation. The larger the circle and the darker the colour, the higher the correlations



Radiomics feature extraction, standardisation, and selection

A total of 1037 radiomics features were extracted (14 shape-based features, 18 first-order statistical features, 75 texture features, and 744 wavelet decomposition features, 93 exponentials,

93 square transform features) from the tumour VOIs. Figure 4a shows that the data distributions of the two centres were relatively scattered before eliminating the centre effects, whereas these were pooled together following normalisation using ComBat.

For feature selection, 590 features were coarsely selected by an independent-samples *t*-test or Mann-Whitney *U* test.

Table 2 Results of univariate and multivariate logistic regression analysis

Clinical factors	Univariate analysis		Multivariate analysis	
	OR (95%CI)	<i>p</i> value	OR (95%CI)	<i>p</i> value
Age	0.972 (0.958–0.985)	< 0.001	0.980 (0.964–0.996)	0.013
Gender (reference: female)				
Male	1.005 (0.702–1.439)	0.977	-	-
Location (reference: left colon)				
Right colon	1.821 (0.135–0.357)	0.008	1.442 (0.884–2.353)	0.142
Rectum	0.221 (0.135–0.357)	< 0.001	0.417 (0.242–0.715)	0.002
CEA	0.998 (0.996–1.000)	0.137	-	-
Ki-67	1.256 (0.520–3.040)	0.612	-	-
Radiomics signature	2.996 (2.454–3.682)	< 0.001	2.405 (1.935–3.003)	< 0.001

Note. *OR* odds ratio, *CI* confidence interval, *CEA* carcinoma embryonic antigen

Subsequently, 47 high-frequency features that appeared more than 15 times in 100 feature subsets were selected using random downsampling and RFE-SVM methods (Fig. 4b). Finally, 10 predictive features were determined using a backward stepwise logistic regression analysis (Fig. 4c). The correlation results between the features are shown in Fig. 4d.

Development and performance of the MSI-related radiomics signature

Ten radiomics features were employed as network input neurones, MSI statuses were employed as network output neurones, and the number of hidden neurones was set in a range of 1 to 9. Before training, MSI statuses were pre-processed by times two and minus one to fit the tangent activation function. The GANN model was built with the following parameters: maximum epochs, 500; generations, 300; and population size, 100. Initial population individuals were randomly generated. The other parameters were set as the default values.

The AUCs with different hidden neurones in the three cohorts are shown as three broken line graphs (Fig. 5a). After considering the overall performance of the different cohorts, the optimal structure was determined to be the GANN with five hidden neurones and AUCs of 0.792 (95% confidence interval [CI]: 0.706–0.873), 0.788 (95% CI: 0.671–0.888), and 0.780 (95% CI: 0.722–0.831), respectively. The detailed optimisation process is shown in Fig. 5b, and the distribution of the network weights and the proportions of radiomics features weights are visualised as a Sankey diagram (Fig. 5c).

Development and performance of the prediction models

Univariate and multivariate logistic analyses identified age, tumour location, and radiomics signature as independent

predictors of MSI status, as detailed in Table 2. The combined model was constructed by combining all the above factors, while the clinical model was constructed with only age and tumour location.

The AUC, accuracy, sensitivity, and specificity of each model are presented in Table 3. The performance of the MSI-related radiomics signature (AUCs of 0.792, 0.788, and 0.775 in the training, internal, and external validation cohorts) was comparable to that of the combined model (AUCs of 0.806, 0.777, and 0.767; DeLong's test, $p = 0.746–0.920$); it also outperformed the clinical model (AUCs of 0.747, 0.768, and 0.623, $p < 0.001$).

Prognostic analysis of the signature

Patients were divided into predicted MSI-L/MSS (signature < -0.01) or MSI-H (signature > -0.01) groups based on the signature cutoff value by maximising their Youden index values. Kaplan–Meier survival analyses (Fig. 6) showed that the signature could successfully stratify stage II CRC patients (HR: 0.402, 95% CI: 0.172–0.937; log-rank, $p = 0.029$) and predicted that patients with MSI-H had a significant prognostic advantage compared to MSI-L/MSS patients, while no statistical differences were found for stage I ($p = 0.940$) and III ($p = 0.950$) CRC patients.

Discussion

In this study, we applied a genetic algorithm–enhanced artificial neural network to establish an MSI-related radiomics signature based on 277 venous phase–enhanced CT images to predict the MSI statuses of CRC patients. Its performance was successfully verified in internal (119 patients) and external validation cohorts (441 patients). Meanwhile, we found that patients with stage II CRC with predicted MSI-H had better

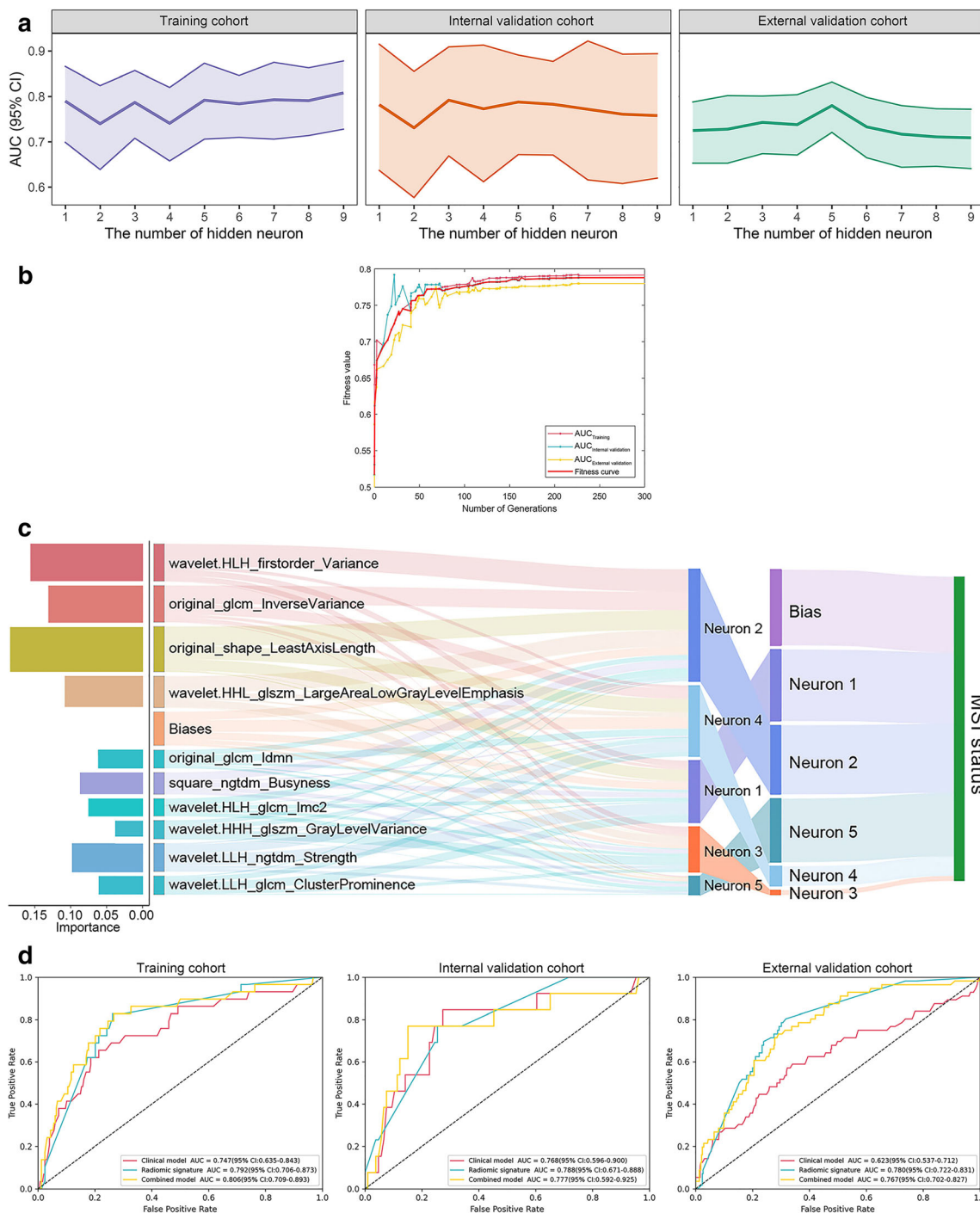


Fig. 5 Model performance analysis. **a** The area under the curve (AUC) broken line graphs of the training, internal, and external validation cohorts, respectively. The bold lines indicate the AUC values, and two thin lines indicate the 95% confidence intervals of the AUCs. **b** Genetic algorithm fitness curve. As generation numbers increased, the individual optimal fitness was continuously improved. The optimal network weight combination was obtained at 227 generations. **c** Sankey diagrams. Left,

each feature’s importance is calculated by the ratio of their respective weight to the total network weight. Right, the relationships among the absolute value of the feature weight, five hidden neurons, and MSI status. **d** Receiver operating characteristic (ROC) curves. The AUC and 95% confidence intervals of the training, internal, and external validation cohorts for the different models are shown in the bottom right

prognoses than those with MSI-L/MSS. The results indicate that the radiomics method is a feasible tool for assessing MSI status.

Although judged by AUC values, the performance of our radiomics signature was moderate compared to previous models [11–18] (AUCs of 0.775–0.792 versus 0.688–0.918,

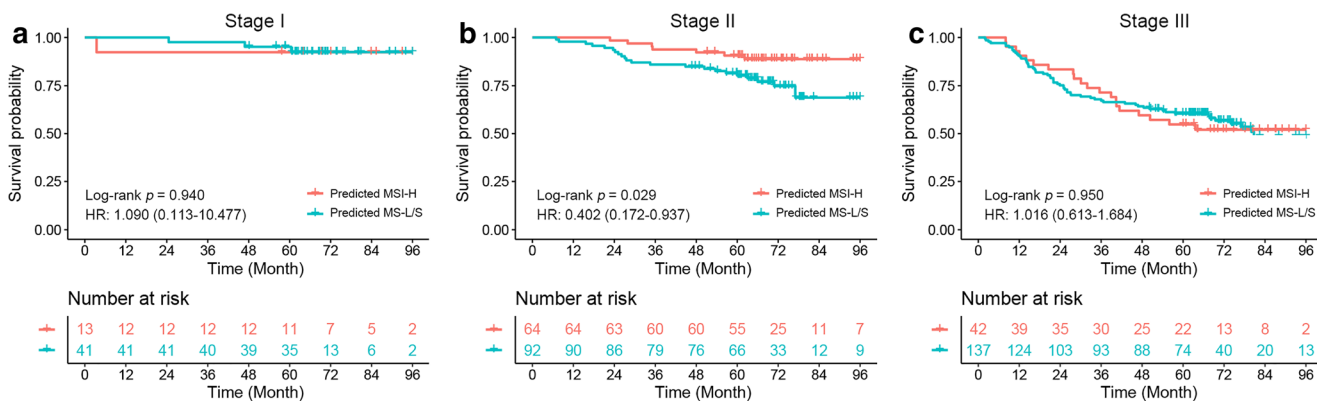
Table 3 Performance of prediction models for MSI status prediction

Cohort	Model	AUC (95% CI)	Accuracy	Sensitivity	Specificity
Training	Clinical model	0.747 (0.644–0.838)	0.773	0.655	0.786
	Radiomics signature	0.792 (0.703–0.872)	0.747	0.828	0.738
	Combined model	0.806 (0.696–0.890)	0.744	0.828	0.734
Internal validation	Clinical model	0.768 (0.591–0.896)	0.756	0.615	0.774
	Radiomics signature	0.788 (0.678–0.899)	0.697	0.769	0.689
	Combined model	0.777 (0.605–0.924)	0.706	0.769	0.698
External validation	Clinical model	0.623 (0.535–0.714)	0.728	0.446	0.769
	Radiomics signature	0.775 (0.720–0.826)	0.726	0.732	0.725
	Combined model	0.767 (0.701–0.828)	0.721	0.714	0.722

respectively); our model had a more robust performance when applied to the real-world data cohorts in our study. This might have resulted from the elimination of potential centre effects with the ComBat method and well-designed feature selection and modelling strategies applied to account for imbalanced data.

Unbalanced data distribution owing to the low incidence of MSI is a challenge for MSI status classification in real-world applications. Some previous studies used the synthetic minority oversampling technique (SMOTE) [16, 18] to create simulated MSI samples; however, this strategy made the model more likely to become overfitted. Another study proposed an initial bias adjustment method [12] for weighted MSI samples, but it was difficult to guarantee the generalisation ability. Bootstrapping is a recommended method for handling imbalanced medical image data [20]. In our study, a combination of bootstrapping and the RFE-SVM method was proposed to select high-frequency features, and an upsampling modelling strategy was adopted to alleviate model bias toward the majority class [29]. Moreover, a genetic algorithm was utilised to further improve the generalisation ability of the signature. This may be attributed to the fact that genetic algorithm is good at global searching, and the weight adjustment is also exquisite [30].

Our results demonstrate the feasibility of using CT radiomics features to predict CRC genetic information. Although the underlying mechanisms by which radiomics predicts MSI status remain unclear, we assumed that radiomics features could characterise tumour size and heterogeneity. Most studies have reported that MSI-H CRC showed a larger tumour size and more intra-tumoural heterogeneity than MSI-L/MSS tumours [31, 32]. Based on the feature weights in the signature, LeastAxisLength, Variance, and InverseVariance were ranked as the top three radiomics features. LeastAxisLength measures the length of the VOI-enclosing ellipsoid and represents the tumour size. Variance and InverseVariance are the measurements of image spatial heterogeneity, and larger values of variance and smaller values of InverseVariance reflect greater heterogeneity. According to the radiomics signature equation, tumours with larger sizes and more heterogeneity tended to be MSI-H, which is partly in agreement with the findings of Chaddad et al [33]. In contrast, MSI status is a prognostic biomarker, and patients with stage II MSI-H CRC have a lower risk of death [5]. Interestingly, we found that patients with stage II predicted MSI-H CRC (HR: 0.402, 95% CI: 0.172–0.937; $p = 0.029$) showed better outcomes than those with predicted MSI-L/MSS.

**Fig. 6** Kaplan–Meier curves according to signature predicted MSI status for colorectal cancer (CRC) patients with stage I (a), II (b), and III (c) tumours

Moreover, clinical factors are associated with genetic alterations in tumours. Consistent with previous studies [34], our study identified that MSI-H CRC occurs predominantly in the younger population and on the right side. This may be explained by the fact that the proximal and distal colon have different embryonic origins, leading to distinct biological properties [35]. However, the addition of clinical factors to radiomics features did not significantly enhance our model's prediction performance, which is in line with previous studies [36, 37]. This finding may be attributed to the fact that radiomics methods can capture heterogeneity differences among tumour locations and different genetic and epigenetic information, a function that overlaps with but is ultimately superior to clinical effects after employing powerful machine learning algorithms [38].

This study had some limitations. First, this was a retrospective study, which may have inevitably led to some information selection bias. Second, we focused only on the most commonly used CT images of the portal venous phase to develop the radiomics signature. Further work is needed to explore whether non-contrast and arterial phase images can provide more information for MSI status estimation. Third, CRC tumours were manually segmented, which is time-consuming and inevitably causes inter-observer variations. In future studies, an automated colorectal tumour segmentation method should be developed to resolve this issue. Finally, different scanners and acquisition parameters were used in our study, which may have influenced the radiomics features. However, significant efforts have been made to minimise the effects using ComBaTool and the *z*-scores method.

In conclusion, we constructed and validated a non-invasive radiomics signature based on routine clinical CT images to identify patients' MSI statuses, and the signature could stratify patients with stage II CRC. These findings could potentially aid in clinical decision-making regarding personalised treatment for patients with CRC.

Funding This study was supported by the Key-Area Research and Development Program of Guangdong Province [No. 2021B0101420006], the National Key Research and Development Program of China [No. 2021YFF1201003], the National Science Fund for Distinguished Young Scholars [No. 81925023], the National Natural Scientific Foundation of China [No. 82072090, 81901910, 81771912, and 81901909], Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application [No. 2022B1212010011], Funding by Science and technology Projects in Guangzhou [No. 202201020001, 202201010513], and High-level Hospital Construction Project [No. DFJHBF202105].

Declarations

Guarantor The scientific guarantor of this publication is Xiaobo Chen.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry Xiaobo Chen, Lan He, and Xin Chen kindly provided statistical advice for this manuscript.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology

- retrospective
- diagnostic or prognostic study
- multicentre study

References

1. Latham A, Srinivasan P, Kemel Y et al (2019) Microsatellite instability is associated with the presence of Lynch syndrome pan-cancer. *J Clin Oncol* 37:286–295
2. Wei Q, Ye Z, Zhong X et al (2017) Multiregion whole-exome sequencing of matched primary and metastatic tumours revealed genomic heterogeneity and suggested polyclonal seeding in colorectal cancer metastasis. *Ann Oncol* 28:2135–2141
3. Nguyen M, Tipping Smith S, Lam M et al (2021) An update on the use of immunotherapy in patients with colorectal cancer. *Expert Rev Gastroenterol Hepatol* 15:291–304
4. Boland CR, Goel A (2010) Microsatellite instability in colorectal cancer. *Gastroenterology* 138(6):2073–2087
5. Toh JWT, Phan K, Reza F, Chapuis P, Spring KJ (2021) Rate of dissemination and prognosis in early and advanced stage colorectal cancer based on microsatellite instability status: systematic review and meta-analysis. *Int J Color Dis* 36:1573–1596
6. Toh JWT, Singh P, Tangirala VAASK, Limmer A, Spring KJ (2021) A simple, accurate and cost-effective capillary electrophoresis test with computational methods to aid in universal microsatellite instability testing. *Cells* 10:1401
7. Yörüker EE, Holdenrieder S, Gezer U (2016) Blood-based biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *Clin Chim Acta* 455:26–32
8. Marcuello M, Vymetalkova V, Neves RPL et al (2019) Circulating biomarkers for early detection and clinical management of colorectal cancer. *Mol Asp Med* 69:107–122
9. Huang Y, Liang C, He L et al (2016) Development and validation of a radiomics nomogram for preoperative prediction of Lymph node metastasis in colorectal cancer. *J Clin Oncol* 34:2157–2164
10. Liu Z, Meng X, Zhang H et al (2020) Predicting distant metastasis and chemotherapy benefit in locally advanced rectal cancer. *Nat Commun* 11:4308
11. Pei Q, Yi X, Chen C et al (2022) Pre-treatment CT-based radiomics nomogram for predicting microsatellite instability status in colorectal cancer. *Eur Radiol* 32:714–724
12. Cao Y, Zhang G, Zhang J et al (2021) Predicting microsatellite instability status in colorectal cancer based on triphasic enhanced computed tomography radiomics signatures: a multicenter study. *Front Oncol* 11:687771
13. Li Z, Zhong Q, Zhang L et al (2021) Computed tomography-based radiomics model to preoperatively predict microsatellite instability status in colorectal cancer: a multicenter study. *Front Oncol* 11:666786
14. Li J, Yang Z, Xin B et al (2021) Quantitative prediction of microsatellite instability in colorectal cancer with preoperative PET/CT-based radiomics. *Front Oncol* 11:702055
15. Wu J, Zhang Q, Zhao Y et al (2019) Radiomics analysis of iodine-based material decomposition images with dual-energy computed

- tomography imaging for preoperatively predicting microsatellite instability status in colorectal cancer. *Front Oncol* 9:1250
16. Li Z, Dai H, Liu Y, Pan F, Yang Y, Zhang M (2021) Radiomics analysis of multi-sequence MR images for predicting microsatellite instability status preoperatively in rectal cancer. *Front Oncol* 11:697497
 17. Golia Pernicka JS, Gagniere J, Chakraborty J et al (2019) Radiomics-based prediction of microsatellite instability in colorectal cancer at initial computed tomography evaluation. *Abdom Radiol (NY)* 44:3755–3763
 18. Fan S, Li X, Cui X et al (2019) Computed tomography-based radiomic features could potentially predict microsatellite instability status in stage II colorectal cancer: a preliminary study. *Acad Radiol* 26:1633–1640
 19. Bader-El-Den M, Teitei E, Perry T (2019) Biased random forest for dealing with the class imbalance problem. *IEEE Trans Neural Netw Learn Syst* 30:2163–2172
 20. Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts HJWL (2018) Data analysis strategies in medical imaging. *Clin Cancer Res* 24:3492–3499
 21. Dissaux G, Visvikis D, Da-Ano R et al (2020) Pretreatment F-FDG PET/CT radiomics predict local recurrence in patients treated with stereotactic body radiotherapy for early-stage non-small cell lung cancer: a multicentric study. *J Nucl Med* 61:814–820
 22. Ligerio M, Jordi-Ollero O, Bernatowicz K et al (2021) Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur Radiol* 31:1460–1470
 23. Ji G-W, Zhu F-P, Xu Q et al (2020) Radiomic features at contrast-enhanced CT predict recurrence in early stage hepatocellular carcinoma: a multi-institutional study. *Radiology* 294:568–579
 24. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 77:e104–e107
 25. Orlhac F, Lecler A, Savatovski J et al (2021) How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol* 31:2272–2280
 26. Da-Ano R, Masson I, Lucia F et al (2020) Performance comparison of modified ComBat for harmonisation of radiomic features for multicenter studies. *Sci Rep* 10:10248
 27. Zhang Y, Gao X, Smith K et al (2019) Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water Res* 164:114888
 28. Gong Y-J, Li J-J, Zhou Y et al (2016) Genetic learning particle swarm optimization. *IEEE Trans Cybern* 46:2277–2290
 29. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell* 5:221–232
 30. Yu L, Jin W, Li X, Zhang Y (2018) Optimization of bioactive ingredient extraction from Chinese herbal medicine : a comparative study of three optimization models. *Evid Based Complement Alternat Med* 2018:6391414
 31. Fujiyoshi K, Yamaguchi T, Kakuta M et al (2017) Predictive model for high-frequency microsatellite instability in colorectal cancer patients over 50 years of age. *Cancer Med* 6:1255–1263
 32. Bai W, Ma J, Liu Y et al (2019) Screening of MSI detection loci and their heterogeneity in East Asian colorectal cancer patients. *Cancer Med* 8:2157–2166
 33. Bogowicz M, Riesterer O, Ikenberg K et al (2017) Computed tomography radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. *Int J Radiat Oncol Biol Phys* 99:921–928
 34. Song Y, Wang L, Ran W et al (2020) Effect of tumor location on clinicopathological and molecular markers in colorectal cancer in Eastern China patients: an analysis of 2,356 cases. *Front Genet* 11:96
 35. Bufill JA (1990) Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann Intern Med* 113:779–788
 36. Han Y, Xie Z, Zang Y et al (2018) Non-invasive genotype prediction of chromosome 1p/19q co-deletion by development and validation of an MRI-based radiomics signature in lower-grade gliomas. *J Neuro-Oncol* 140:297–306
 37. Gong L, Xu M, Fang M et al (2020) Noninvasive prediction of high-grade prostate cancer via biparametric MRI radiomics. *J Magn Reson Imaging* 52:1102–1109
 38. Sagaert X, Vanstapel A, Verbeek S (2018) Tumor heterogeneity in colorectal cancer: what do we know so far? *Pathobiology* 85:72–84

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