Development and interpretation of a clinicopathological-based model for the identification of microsatellite instability in Colorectal Cancer

Zhenxing Jiang
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Yinghao Cao
Department of Digestive Surgical Oncology, Cancer Center, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430022, People's Republic of China.

Lizhao Yan
Department of Hand Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China;

Shenghe Deng
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Junnan Gu
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Le Qin
Department of General Surgery, First Affiliated Hospital, School of Medicine, Shihezi University, Shihezi, Xinjiang 832008, P.R. China.

Fuwei Mao
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Yifan Xue
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Wentai Cai
College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Xiu Nie
Department of Pathology, Union Hospital, Tongji Medical, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Hongli Liu
Cancer Center, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China.
Fumei Shang
Department of Medical Oncology, Nanyang Central Hospital, Nanyang, Henan, China
Kailin Cai (caikailin@hust.edu.cn)
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.
Kaixiong Tao
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.
Jiliang Wang
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.
Ke Wu
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430022, China.

Research Article

Keywords: colorectal cancer, mismatch repair, microsatellite instability, machine learning, predictive model.

Posted Date: June 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1662236/v1

Abstract

Background:

Chemotherapy is not recommended for patients with deficient mismatch repair(dMMR)in colorectal cancer(CRC), so assessing the status of MMR is crucial for the selection of subsequent treatment. Therefore, this study aims to build predictive models to accurately and rapidly identify dMMR.

Methods:

A retrospective analysis based on the clinicopathological data of patients with colorectal cancer from May 2017 to December 2019 at Wuhan Union Hospital was performed. The variables were subjected to co-linearity analysis, least absolute shrinkage and selection operator (LASSO) regression and random forest (RF) feature screening. Four sets of machine learning models (i.e., extreme gradient boosting (XGBoost), support vector machine (SVM), naive bayes (NB), and random forest and a conventional logistic regression (LR) model were built for training and testing. Receiver operating characteristic (ROC) curves are plotted to evaluate the predictive performance.

Findings:

A total of 2279 patients were included in the analysis and were randomly divided into a training group and a test group.The final 12 clinicopathological features were incorporated into the predictive models. The Area Under the curve (AUC) values of the five predictive models were 0.8931 in XGBoost, 0.8906 in SVM, 0.8512 in NB, 0.9088 in RF and 0.8319 in LR. The results showed that the random forest exhibited the best recognition ability and outperformed the conventional logistic regression method in identifing dMMR and proficient mismatch repair (pMMR).

Conclusion:

Our predictive models built on routine clinicopathological data can significantly improve the diagnostic performance of dMMR and pMMR. Meanwhile, four machine learning models outperformed conventional logistic regression model.

1. Introduction

Colorectal cancer is one of the most common cancers in the world and the second leading cause of cancer deaths.(1) Deficient Mismatch Repair (dMMR) presents in 10–20% of colorectal cancer (CRC) patients, suggesting that CRC with dMMR is a biologically distinct type with broad prognostic, predictive and therapeutic importance.(2) Furthermore, DNA mismatch repair system is an evolved and conserved process for repairing errors during replication of proliferating cells.(3) Molecularly targeted therapies and chemotherapeutic agents are used to treat patients with dMMR colorectal cancer.(4) Recently, a growing body of evidence suggests that the individual treatment response of CRC patients is strongly related to its molecular characteristics.(5)

Microsatellite instability (MSI) is the abnormal shortening or lengthening of 1–6 repeat base pair units of DNA, which is caused by inactivation of the DNA MMR system.(3, 6, 7) Colorectal cancer patients with microsatellite instability are more likely to find Lynch syndrome.(8, 9) Thus, MMR is essential to ensure genetic information stability and avoid future genetic diseases.(10) According to the National Comprehensive Cancer Network (NCCN), patients with stage II colorectal cancer with MSI or dMMR are simply observed after surgery and do not require chemotherapy, which is fortunate for many colorectal cancer patients. A study by Klingbiel D(11) and Michael J. Overman(12) showed that colorectal cancer patients with MSI are insensitive to pentafluorouracil chemotherapy, but sensitive to PD-1 immunotherapy, which provides more rationalization of colorectal cancer treatment. However, most patients are unable to undergo genetic testing to detect dMMR status due to the cost of money and time.

Recently, artificial intelligence has become a research hotspot in medicine, with the promise of achieving a high-precision automated diagnosis of heterogeneous diseases. ole-Johan Skrede et al.(13) utilized deep learning combined with conventional digital scanning of hematoxylin and eosin-stained tumour tissue sections to develop a clinically useful prognostic marker that can classify stage II and III patients into different prognostic groups and then guide the application of adjuvant chemotherapy. Frederick Matthew Howard et al.(14) used a machine learning model to successfully predict which patients who underwent surgery should remove squamous cell carcinoma of the neck and who were at intermediate risk would benefit from receiving cisplatin-based chemoradiation therapy (CRT). Quirino La et al.(15) found higher accuracy in predicting survival after liver cancer treatment using artificial intelligence compared to traditional linear analysis systems. In addition, Yu et al.(16) showed that they applied seven machine learning classifiers to predict survival time of lung cancer patients based on histopathological features and obtained a fairly satisfactory prediction accuracy. However, no study cant systematically evaluate the detection value of machine learning models based on simple clinicopathological indicators in dMMR. Therefore, a simple, minimally invasive and accurate method to identify dMMR is urgently needed.

Based on simple clinicopathological indicators and previous studies, this study developed four machine learning models and a logistic regression model to predict colorectal cancer lacking DNA mismatch repair, thus helping clinicians to identify MMR status and providing a reference for precise treatment plan for patients.

2. Method

2.1. Study population

Retrospective analysis of 2279 colorectal cancer patients treated for confirmed disease at Wuhan Union Hospital from May 2017 to December 2019. Patients with the following conditions were excluded from the study: i) no MMR status outcome; ii) no complete clinical data; and iii) history of radiotherapy and chemotherapy prior to MMR status identification. A total of 2279 patients were enrolled in our study and randomly assigned to the training and test sets in a 7-to-3 ratio. The detailed process of patient selection

is shown in Fig.1. The study protocol was reviewed and approved by the Ethics Committee and institutional Review Committee of Wuhan Union Hospital(No.2018-S377). This study was performed in line with the principles of the Declaration of Helsinki. All patients signed an informed consent form stating that they understood the procedure and its potential complications and agreed to participate in this study.

2.2. Data Collection

Baseline clinicopathological information on the patients obtained from the hospital's medical records included serum tumour markers - carcinoembryonic antigen (CEA), glycoantigen 19-9 (CA19-9), glycoantigen 12-5 (CA12-5), glycoantigen 72-4 (CA72-4), glycoantigen 15-3 (CA15-3), alpha-fetoprotein (AFP), serum squamous cell carcinoma antigen (SCC), ferritin (FERR), cytokeratin 19 fragment cyfra21-1 (CYFRA21-1), serum neuron-specific enolase (NSE), pathological type, histological type, age, sex, location, diameters, number of sampled lymph nodes (LNs), number of positive LNs, T stage, N stage, M stage, perineural invasion and vascular invasion. MMR status was assessed by immunohistochemistry (IHC) and was determined by MSH2, MSH6, MLH1 and PMS2 markers. We defined dMMR as a lacking expression of one or more MMR proteins, while tumour with intact MMR proteins was categorized as pMMR.

2.3. Four machine learning classifiers and a conventional logistic regression model

In this study, we built four machine learning models (i.e., XGBoost, SVM, NB, and RF) and a conventional LR model using the caret package for R language (version 6.0-90) to diagnose dMMR discriminatively. An analysis of co-linearity was performed on the initial 23 variables to exclude significantly correlated variables. Subsequently, LASSO regression and RF were used for variable selection. This process utilized five times ten-fold cross-validation to ensure the reliability of the results. The data were randomly divided into training and validation sets by 7:3. The variables screened by LASSO regression and random forest methods were integrated and incorporated into the predictive models. Ten-fold cross-validation and 10*10 grid research were used for model hyperparameter selection.

2.4. Data analysis

Continuous variables between the dMMR and pMMR groups were analysed using the Student t-test or the Mann-Whitney U test (as appropriate). Also, categorical data were compared with the chi-square test or Fisher's exact test. Receiver operating characteristic (ROC) curve were performed to assess the diagnostic performance of predictive models of dMMR. The area under the curve (AUC) was measured in each ROC curve, and specificity and sensitivity were calculated to assess the diagnostic performance of five models. The above statistical analyses were performed using the R software version. Differences were considered statistically significant when $P<0.05$ for both sides.

3. Results

3.1. Patient Characteristics

All 2279 colorectal cancer patients from Wuhan Union Hospital have complete clinical information and are finally selected for this study. Of the 2279 colorectal patients, 177 were diagnosed with dMMR (7.77%). The consensus criteria for dMMR protein diagnosis was to select CRC patients who met the Revised Bethesda Guidelines (RBG) and then underwent MSI testing and/or immunohistochemical staining for MMR protein. The differences in clinicopathological characteristics between the dMMR and pMMR groups in the Wuhan union hospital (WUH) cohort are shown in Table 1. We could observe statistical differences between the dMMR and pMMR groups in terms of age, primary site, tumour diameter, histology, number of lymph nodes, number of positive lymph nodes, N stage, TNM stage, peripheral perineural invasion, ferritin and some serum tumour markers. The internal categorical distribution of each variable is shown in Supplementary Fig. 1.

Table 1 Clinical characteristics of the patients with colorectal cancer.

	Level	Overall	dMMR	pMMR	P-value
n		2279	177	2102	
Gender (%)	Male	1369 (60.1)	107 (60.5)	1262 (60.0)	0.978
	Female	910 (39.9)	70 (39.5)	840 (40.0)	
Age (%)	<53	833 (36.6)	89 (50.3)	744 (35.4)	**<0.001**
	>=53	1446 (63.4)	88 (49.7)	1358 (64.6)	
Primary location (%)	colon	1082 (47.5)	159 (89.8)	923 (43.9)	**<0.001**
	rectum	1197 (52.5)	18 (10.2)	1179 (56.1)	
Tumor diameters (cm) (%)	<4.6	1420 (62.3)	52 (29.4)	1368 (65.1)	**<0.001**
	>=4.6	859 (37.7)	125 (70.6)	734 (34.9)	
Pathological type (%)	non-adenocarcinoma	576 (25.3)	61 (34.5)	515 (24.5)	**0.005**
	adenocarcinoma	1703 (74.7)	116 (65.5)	1587 (75.5)	
Histology (%)	Well/moderate	1976 (86.7)	137 (77.4)	1839 (87.5)	**<0.001**
	poor	303 (13.3)	40 (22.6)	263 (12.5)	
No of sampled LNs (n) (%)	<23	1735 (76.1)	85 (48.0)	1650 (78.5)	**<0.001**
	>=23	544 (23.9)	92 (52.0)	452 (21.5)	
No of Positive LNs (n) (mean (SD))		2.02 (3.85)	0.89 (2.90)	2.12 (3.90)	**<0.001**
T-stage (%)	I/II	405 (17.8)	20 (11.3)	385 (18.3)	**0.025**
	III/IV	1874 (82.2)	157 (88.7)	1717 (81.7)	
N-stage (%)	N0	1249 (54.8)	134 (75.7)	1115 (53.0)	**<0.001**

	Level	Overall	dMMR	pMMR	P-value
	N2	430 (18.9)	11 (6.2)	419 (19.9)	
	N1	600 (26.3)	32 (18.1)	568 (27.0)	
M-stage (%)	0	2237 (98.2)	174 (98.3)	2063 (98.1)	1.000
	1	42 (1.8)	3 (1.7)	39 (1.9)	
TNM (%)	1	319 (14.0)	18 (10.2)	301 (14.3)	**<0.001**
	2	913 (40.1)	115 (65.0)	798 (38.0)	
	3	1005 (44.1)	41 (23.2)	964 (45.9)	
	4	42 (1.8)	3 (1.7)	39 (1.9)	
Perineural invasion (%)	No	1549 (68.0)	160 (90.4)	1389 (66.1)	**<0.001**
	Yes	730 (32.0)	17 (9.6)	713 (33.9)	
vascular cancer embolus (%)	No	1734 (76.1)	146 (82.5)	1588 (75.5)	**0.047**
	Yes	545 (23.9)	31 (17.5)	514 (24.5)	
CEA (%)	Normal	1357 (59.5)	122 (68.9)	1235 (58.8)	**0.010**
	High	922 (40.5)	55 (31.1)	867 (41.2)	
CA72-4 (%)	Normal	1891 (83.0)	120 (67.8)	1771 (84.3)	**<0.001**
	High	388 (17.0)	57 (32.2)	331 (15.7)	
CA199 (%)	Normal	1855 (81.4)	144 (81.4)	1711 (81.4)	1.000
	High	424 (18.6)	33 (18.6)	391 (18.6)	
AFP (%)	Low	298 (13.1)	34 (19.2)	264 (12.6)	**0.036**

	Level	Overall	dMMR	pMMR	P-value
	Normal	1957 (85.9)	142 (80.2)	1815 (86.3)	
	High	24 (1.1)	1 (0.6)	23 (1.1)	
SCC (%)	Normal	2174 (95.4)	168 (94.9)	2006 (95.4)	0.897
	High	105 (4.6)	9 (5.1)	96 (4.6)	
NSE (%)	Normal	1513 (66.4)	122 (68.9)	1391 (66.2)	0.508
	High	766 (33.6)	55 (31.1)	711 (33.8)	
CA125 (%)	Normal	2060 (90.4)	157 (88.7)	1903 (90.5)	0.508
	High	219 (9.6)	20 (11.3)	199 (9.5)	
CA15-3 (%)	Normal	872 (38.3)	83 (46.9)	789 (37.5)	**0.017**
	High	1407 (61.7)	94 (53.1)	1313 (62.5)	
FERR (%)	Low	1134 (49.8)	123 (69.5)	1011 (48.1)	**<0.001**
	Normal	1018 (44.7)	49 (27.7)	969 (46.1)	
	High	127 (5.6)	5 (2.8)	122 (5.8)	
CYFRA21-1 (%)	Normal	1706 (74.9)	132 (74.6)	1574 (74.9)	1.000
	High	573 (25.1)	45 (25.4)	528 (25.1)	

3.2. Construction of predictive Models

Twenty-three variables were initially included based on simple clinicopathological data of the patients. Co-linearity between variables was excluded before modelling. The results of the variable correlation analysis (Fig. 2) showed that there was no co-linearity among independent variables. To make the model more practical and simple, we further selected the initial 23 variables using LASSO regression and random forest, and the selecting results are shown in Figs. 3 and 4. The λ value of binomial deviation under one standard error was used for the final LASSO regression by performing five times ten-fold cross-

validation method. The LASSO regression and random forest were selected for 9 and 11 variables, respectively. The process associated with variable selecting is also shown in Supplementary Fig. 3 and Fig. 4. Meanwhile, we combined the variables screened by both methods. The final 12 variables were included in the predictive models. And, twelve clinicopathological characteristics were used as the best subset of risk factors as the final parameters for model input (Table2).

Table 2 Risk factors for deficient MMR in Colorectal Cancer.

Characteristic	OR[1]	95% CI[1]	p-value
Age			**<0.001**
<53	—	—	
>=53	2.93	1.54, 5.72	
Primary.location			**<0.001**
colon	—	—	
rectum	10.5	4.95, 23.9	
Tumor.diameters..cm.			**<0.001**
<4.6	—	—	
>=4.6	0.24	0.12, 0.46	
Pathological.type			0.23
non-adenocarcinoma	—	—	
adenocarcinoma	1.53	0.76, 3.10	
Histology			0.21
Well/moderate	—	—	
poor	0.59	0.25, 1.35	
No..of.sampled.LNs..n.			**0.009**
<23	—	—	
>=23	0.42	0.21, 0.80	
N.stage			0.76
N0	—	—	
N1	2.67	0.02, 515	
N2	5.58	0.02, 1,683	
Perineural.invasion			**0.028**
No	—	—	
Yes	2.87	1.12, 7.85	
NSE			**0.018**
Normal	—	—	
High	2.20	1.15, 4.33	

Characteristic	OR[1]	95% CI[1]	p-value
No..of.Positive.LNs..n.	1.14	0.87, 1.73	0.47
CA72.4			0.23
Normal	—	—	
High	0.65	0.31, 1.32	
TNM			0.71
1	—	—	
2	1.86	0.65, 5.34	
3	1.30	0.01, 195	
4	1.24	0.02, 90.2	

[1]OR = Odds Ratio, CI = Confidence Interval

3.3. Performance of Models

The 2279 colorectal cancer patients were randomly divided into training and test sets in a 7-to-3 ratio. ROC curve was used to evaluate the performance of the four machine learning models and a LR model. As shown in Fig. 5, the AUC of the test set was as follows: XGBoost was 0.8931, SVM was 0.8906, NB was 0.8512, and RF was 0.9088 and LR was 0.8319. Therefore, we can conclude that the RF model has an excellent predictive ability to identify colorectal cancer with dMMR with a sensitivity of 0.8679 and a specificity of 0.6962. Moreover, machine learning models have a better predictive ability, than to conventional LR method. In addition, the accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the predictive models on the test set are listed in Table 3.

Table 3 Performance of different predictive models to identify dMMR.

Model	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Accuracy	AUC-ROC
LR	0.7170	0.7595	0.6667	0.8000	0.7424	0.7835
RF	0.8679	0.6962	0.6571	0.8871	0.7652	0.8584
NB	0.7547	0.6329	0.5797	0.7937	0.6818	0.7424
SVM	0.7736	0.6709	0.6119	0.8154	0.7121	0.8174
XGboost	0.7170	0.7468	0.6552	0.7973	0.7348	0.8055

3.4. Variable importance analysis

We performed feature importance analysis for the variables seleced by LASSO regression and random forest, respectively, and the results are displayed in Fig. 6 and Fig. 7. The final 12 variables were incorporated into the predictive models for training and validation. To investigate the potential impact of each clinical feature on the predictive model recognition ability, we ranked the clinical variables that showed the best results in the random forest model in order of their contribution to the output results from highest to lowest, as shown in Fig. 8. We found that the top two rankings were the location and diameter of tumour, the same results as in Fig. 6 and Fig. 7. Moreover, 159 (89.8%) colorectal patients with deficient mismatch repair had tumour location in the colon and 18 (10.2%) in the rectum. 52 (29.4%) colorectal patients with mismatch repair deficiency had tumour diameter < 4.6 cm, 125 (70.6%) ≥ 4.6 cm, as shown in table 1. Together, these results suggest that colorectal cancer with deficient mismatch repair is mainly associated with the location and diameter of tumour.

4. Discussion

CRC remains a major health burden with a high mortality rate worldwide.(17) MMR plays a key role in the progression and prognosis of colorectal cancer disease. The latest guidelines recommend chemotherapy for patients with stage II CRC with pMMR even without high-risk factors.(18) With the rapid advancement of medical science in today's world, genetic testing techniques applied to MMR status can be of great help in the individual treatment and management of colorectal cancer patients, but the diagnosis rate of MMR status is still not high.(19–21) There are many studies on mismatch repair in colorectal cancer, but no studies have been conducted to build machine learning models to predict the deficient mismatch repair based on simple clinicopathological indicators. We have already built a simple model to predict the mismatch repair status of colorectal cancer patients based on clinicopathological parameters and tumour markers with good results in the previous phase.(5) This time, we will further focus on whether the machine learning approach is more effective than the conventional predictive model by constructing four machine learning models and a traditional logistic regression based on simple clinicopathological indicators.

Prediction of colorectal cancer patients with dMMR and / or MSI has also been reported. Ms. Amelie Echle(2) and his colleagues applied deep learning to distinguish dMMR from pMMR. The AUC values for the cross-validation cohort, the external validation cohort and the image-processed external validation cohort were 0.92 (0.91–0.93), 0.95 (0.92–0.96) and 0.96 (0.93–0.98). However, their study did not set appropriate underlying true labels, which would be confounded by noisy labels. Cao Rui et al.(22) built a deep learning model based on histopathology to predict MSI. The AUC value for the test set was 0.8848 and the AUC value for the external validation set was 0.8504. However, the sample size of this study was small and further expansion of the training data is needed to improve the model's accuracy. The above two studies were based on pathology section images to predict MSI or dMMR, which is not a simple process. In our study, we established predictive models to predict the status of mismatch repair of colorectal cancer patients based on simple clinicopathological indicators of the patients, which are very easy to obtain, and selected by machine learning and logistic regression. Therefore, our research has important clinical significance.

The application of machine learning models based on the random forest algorithm is common. liu et al. (23) constructed a machine learning to predict whether the quality of life of thyroid cancer patients will decrease in 3 months after surgery. The AUC value was 0.834 and 0.897 for the training and validation groups, respectively. Sergio Grosu et al.(24) built a random forest model based on CT information to distinguish benign and malignant polyps. In the external validation group, the AUC value of the random forest model was 0.91, with a sensitivity of 82% and specificity of 85%. Manabu Takamatsu et al.(25) used machine learning of digital slide images to predict lymph node metastasis in stage T1 colorectal cancer. The area under the ROC curve of the random forest algorithm was 0.938, and the sensitivity and specificity of the optimal threshold were 80.0% and 94.5%, respectively. The machine learning approach was not significantly different from the conventional histological assessment by hematoxylin staining, but had lower false negative cases. Pushpanjali Gupta et al.(26) developed a machine learning model based on the random forest algorithm to predict five-year disease-free survival in patients with colon cancer with an accuracy of 84% and an AUC value of 0.82 ± 0.10. There are many more studies on machine learning models based on random forests.(27–29) The AUC values of five predictive models in our study are 0.8931 for XGBoost, 0.8906 for SVM, 0.8512 for NB, and 0.9088 for RF, 0.8319 for LR. The RF showed the best predictive results.

Our study still has some shortcomings. First, the population in our cohort was from just one region of China (Wuhan), which may limit the generalizability of the predictive models and requires further validation in patients from different geographic regions. Second, this was a non-randomized retrospective analysis. Therefore, there are potential biased comparisons, such as the inclusion of patients and sample selection bias. Finally, our study has only internal validation and further external validation groups are needed to verify the predictive effect. At the same time, our study is still very significantly. To the best of our knowledge, this study is the first to propose a machine learning approach to analyse and model the MMR status of patients based on their simple clinicopathology and tumour markers. Finally, our single-centre sample is large enough so that the conclusions drawn have some reference value.

Conclusion

In this study, we built four sets of machine learning models and a conventional logistic regression model to predict colorectal cancer patients lacking DNA mismatch repair based on simple clinicopathological indicators. Our results show that predictive behaviour can be made accurately and consistently by building machine learning models. At the same time, machine learning models have better performance in identifying dMMR than to conventional logistic regression method. This provides clinicians with important information, reduces the cost of detection, and avoids wasting medical resources.(30) In the future, we will increase the sample size and sample diversity (geographic diversity) for machine learning models. Adding external validation groups and enriching statistical methods to evaluate will increase the prediction performance of the models.

Abbreviations

dMMR:deficient mismatch repair

pMMR:proficient mismatch repair

CRC:colorectal cancer

LASSO:least absolute shrinkage and selection operator

RF:random forest

XGBoost:extreme gradient boosting

SVM:support vector machine

NB:naive bayes

LR:logistic regression

ROC;Receiver operating characteristic

AUC:Area Under the curve

MSI:Microsatellite instability

NCCN:the National Comprehensive Cancer Network

CRT:chemoradiation therapy

CEA:carcinoembryonic antigen

CA:glycoantigen

AFP:alpha-fetoprotein

SCC:squamous cell carcinoma

FERR:ferritin

CYFRA21:1-cytokeratin 19 fragment cyfra21-1

NSE:neuron specific enolase

LNs:lymph nodes

IHC:immunohistochemistry

RBG:Revised Bethesda Guidelines

WUH:Wuhan union hospital

PPV:positive predictive value

NPV:negative predictive value

Declarations

Ethics approval and consent to participate

This study was performed in line with the principles of the Declaration of Helsinki. Studies involving human participants were reviewed and approved by the Ethics Committee and the Institutional Review Committee of Wuhan Union Medical College(No.2018-S377). Informed consent was obtained from all individual participants included in the study. The recruited volunteers were requested to sign an informed consent form.

Consent for publication

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This study was supported by the National Natural Science Foundation of China (No.82170678) and Hubei Province Key Research and Development Program of China (Science and Technology Innovation Special Project No. 2021BAA04 4), and Wuhan Strong Magnetic Field Interdisciplinary Fund (No. WHMF202113). The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing this manuscript.

Authors' contributions

Conceptualisation: Zhenxing Jiang, Yinghao Cao, Lizhao Yan, Shenghe Deng. Acquisition of data, analysis and interpretation of data: Junnan Gu, Le Qin, Fuwei Mao, Yifan Xue, Fumei Shang, Wentai Cai. Writing-original draft: Zhenxing Jiang, Shenghe Deng, Junnan Gu, Le Qin. Writing-review and editing: All authors. Supervision: Ke Wu, Kailin Cai, Xiu Nie, Hongli Liu, Kaixiong Tao, Jiliang Wang.

Acknowledgements

Thank Lizhao Yan for the support of statistical analysis.

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: a cancer journal for clinicians. 2018;68(6):394–424.
2. Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology. 2020;159(4):1406–16 e11.
3. Cohen R, Buhard O, Cervera P, Hain E, Dumont S, Bardier A, et al. Clinical and molecular characterisation of hereditary and sporadic metastatic colorectal cancers harbouring microsatellite instability/DNA mismatch repair deficiency. Eur J Cancer. 2017;86:266–74.
4. Picco G, Cattaneo CM, van Vliet EJ, Crisafulli G, Rospo G, Consonni S, et al. Werner Helicase Is a Synthetic-Lethal Vulnerability in Mismatch Repair-Deficient Colorectal Cancer Refractory to Targeted Therapies, Chemotherapy, and Immunotherapy. Cancer Discov. 2021;11(8):1923–37.
5. Cao Y, Peng T, Li H, Yang M, Wu L, Zhou Z, et al. Development and validation of MMR prediction model based on simplified clinicopathological features and serum tumour markers. EBioMedicine. 2020;61:103060.
6. Hasan S, Renz P, Wegner RE, Finley G, Raj M, Monga D, et al. Microsatellite Instability (MSI) as an Independent Predictor of Pathologic Complete Response (PCR) in Locally Advanced Rectal Cancer A National Cancer Database (NCDB) Analysis. Annals of surgery. 2020;271(4):716–23.
7. O'Malley DM, Bariani GM, Cassier PA, Marabelle A, Hansen AR, De Jesus Acosta A, et al. Pembrolizumab in Patients With Microsatellite Instability-High Advanced Endometrial Cancer: Results From the KEYNOTE-158 Study. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2022;40(7):752–61.
8. Snowsill T, Coelho H, Huxley N, Jones-Hughes T, Briscoe S, Frayling IM, et al. Molecular testing for Lynch syndrome in people with colorectal cancer: systematic reviews and economic evaluation. Health Technol Asses. 2017;21(51):1-+.
9. Gebert J, Gelincik O, Oezcan-Wahlbrink M. Recurrent Frameshift Neoantigen Vaccine Elicits Protective Immunity With Reduced Tumor Burden and Improved Overall Survival in a Lynch Syndrome Mouse Model (vol 161, pg 1288, 2021). Gastroenterology. 2021;161(6):2070-.
10. Amin A, Farrukh A, Murali C, Soleimani A, Praz F, Graziani G, et al. Saffron and Its Major Ingredients' Effect on Colon Cancer Cells with Mismatch Repair Deficiency and Microsatellite Instability. Molecules. 2021;26(13).
11. Klingbiel D, Saridaki Z, Roth AD, Bosman FT, Delorenzi M, Tejpar S. Prognosis of stage II and III colon cancer treated with adjuvant 5-fluorouracil or FOLFIRI in relation to microsatellite status: results of

the PETACC-3 trial. Ann Oncol. 2015;26(1):126–32.

12. Overman MJ, McDermott R, Leach JL, Lonardi S, Lenz HJ, Morse MA, et al. Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. Lancet Oncol. 2017;18(9):1182–91.
13. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestol K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet. 2020;395(10221):350–60.
14. Howard FM, Kochanny S, Koshy M, Spiotto M, Pearson AT. Machine Learning-Guided Adjuvant Treatment of Head and Neck Cancer. JAMA Netw Open. 2020;3(11):e2025881.
15. Lai Q, Spoletini G, Mennini G, Laureiro ZL, Tsilimigras DI, Pawlik TM, et al. Prognostic role of artificial intelligence among patients with hepatocellular cancer: A systematic review. World J Gastroenterol. 2020;26(42):6679–88.
16. Yu KH, Zhang C, Berry GJ, Altman RB, Re C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun. 2016;7:12474.
17. Kanth P, Inadomi JM. Screening and prevention of colorectal cancer. BMJ (Clinical research ed). 2021;374:n1855.
18. Wang FH, Zhang XT, Li YF, Tang L, Qu XJ, Ying JE, et al. The Chinese Society of Clinical Oncology (CSCO): Clinical guidelines for the diagnosis and treatment of gastric cancer, 2021. Cancer Commun. 2021;41(8):747–95.
19. Noll A, P JP, Zhou M, Weber TK, Ahnen D, Wu XC, et al. Barriers to Lynch Syndrome Testing and Preoperative Result Availability in Early-onset Colorectal Cancer: A National Physician Survey Study. Clin Transl Gastroenterol. 2018;9(9):185.
20. Cenin DR, Naber SK, Lansdorp-Vogelaar I, Jenkins MA, Buchanan DD, Preen DB, et al. Costs and outcomes of Lynch syndrome screening in the Australian colorectal cancer population. J Gastroenterol Hepatol. 2018;33(10):1737–44.
21. Eriksson J, Amonkar M, Al-Jassar G, Lambert J, Malmenas M, Chase M, et al. Mismatch Repair/Microsatellite Instability Testing Practices among US Physicians Treating Patients with Advanced/Metastatic Colorectal Cancer. J Clin Med. 2019;8(4).
22. Cao R, Yang F, Ma SC, Liu L, Zhao Y, Li Y, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. Theranostics. 2020;10(24):11080–91.
23. Liu YH, Jin J, Liu YJ. Machine learning-based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy. Support Care Cancer. 2022;30(3):2507–13.
24. Grosu S, Wesp P, Graser A, Maurus S, Schulz C, Knosel T, et al. Machine Learning-based Differentiation of Benign and Premalignant Colorectal Polyps Detected with CT Colonography in an Asymptomatic Screening Population: A Proof-of-Concept Study. Radiology. 2021;299(2):326–35.
25. Takamatsu M, Yamamoto N, Kawachi H, Chino A, Saito S, Ueno M, et al. Prediction of early colorectal cancer metastasis by machine learning using digital slide images. Comput Methods Programs Biomed. 2019;178:155–61.

26. Gupta P, Chiang SF, Sahoo PK, Mohapatra SK, You JF, Onthoni DD, et al. Prediction of Colon Cancer Stages and Survival Period with Machine Learning Approach. Cancers (Basel). 2019;11(12).
27. Schperberg AV, Boichard A, Tsigelny IF, Richard SB, Kurzrock R. Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials. Int J Cancer. 2020;147(9):2537–49.
28. Topcuoglu BD, Lesniak NA, Ruffin MTt, Wiens J, Schloss PD. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. mBio. 2020;11(3).
29. Li Y, Nowak CM, Pham U, Nguyen K, Bleris L. Cell morphology-based machine learning models for human cell state classification. NPJ Syst Biol Appl. 2021;7(1):23.
30. Kacew AJ, Strohbehn GW, Saulsberry L, Laiteerapong N, Cipriani NA, Kather JN, et al. Artificial Intelligence Can Cut Costs While Maintaining Accuracy in Colorectal Cancer Genotyping. Front Oncol. 2021;11:630953.

Figures

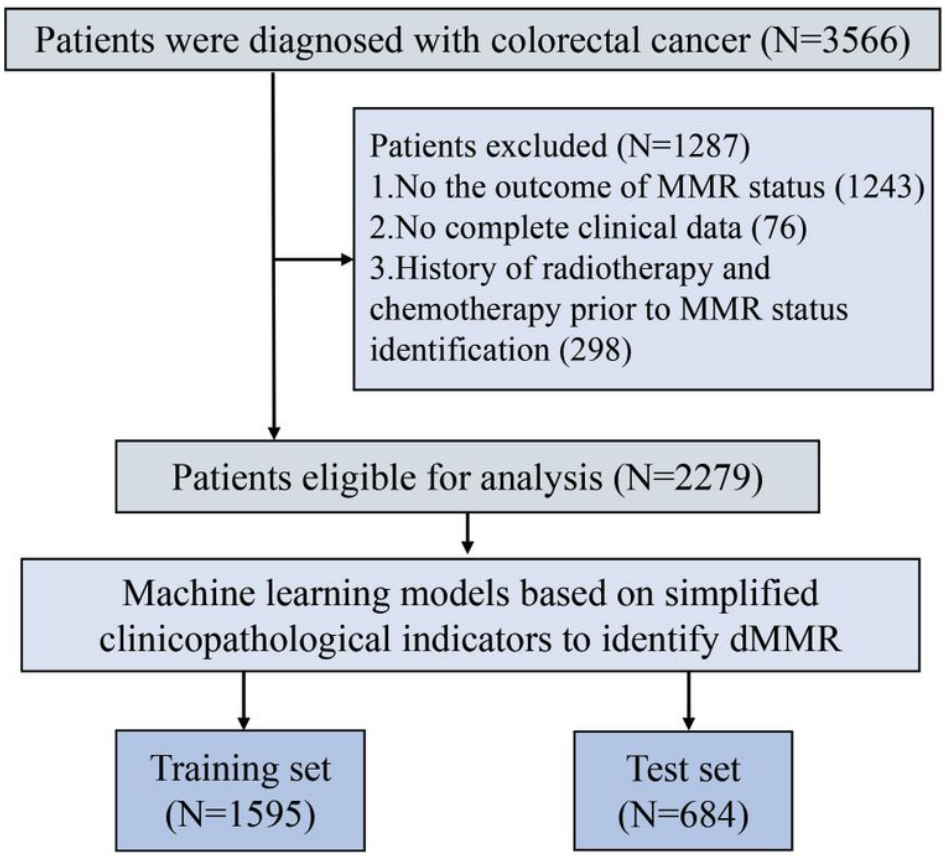

Figure 1

Patient Screening Process

The detailed process of patient selection.

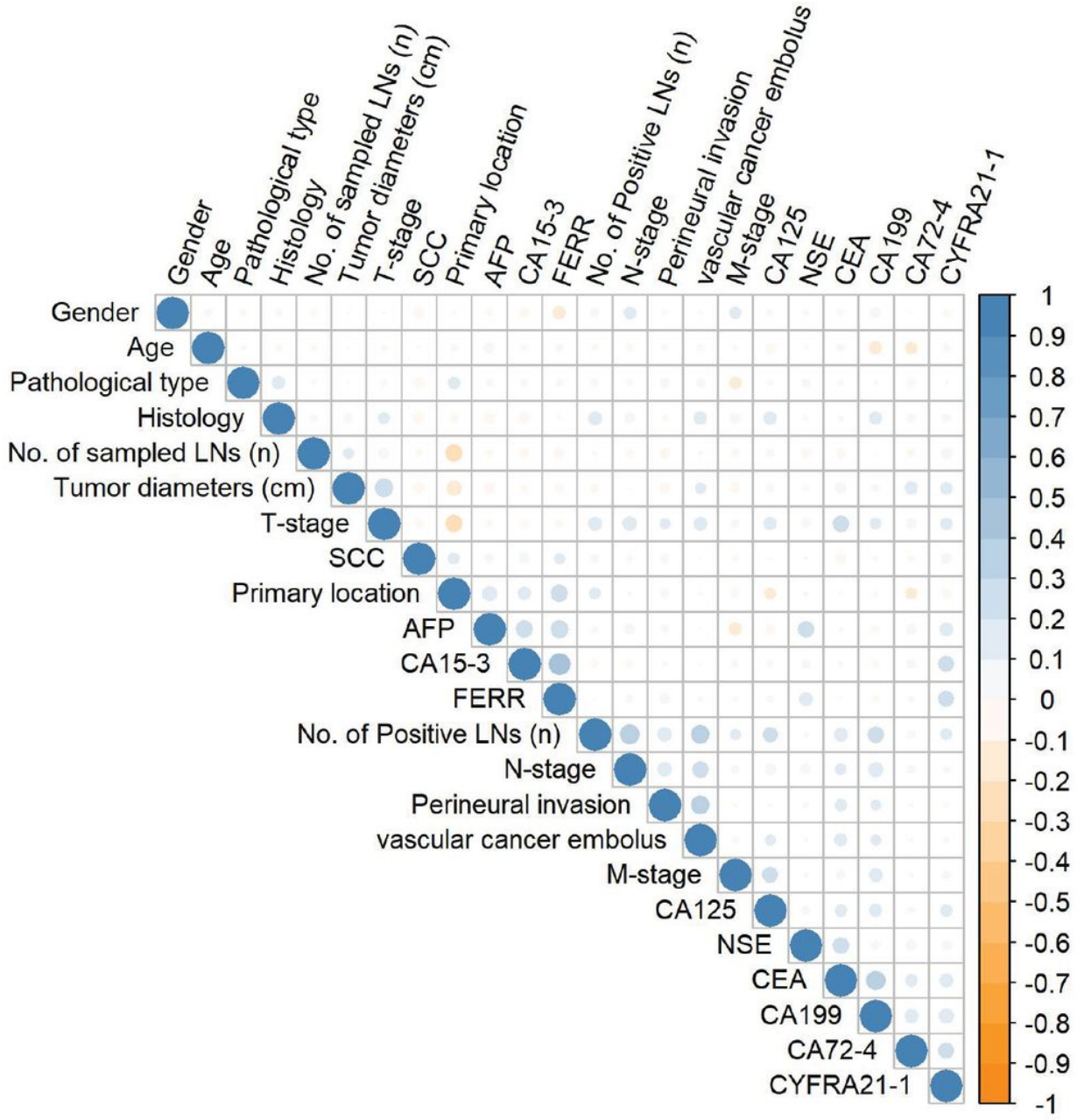

Figure 2

co-linearity analysis

Variables exhibiting *co-linearity* were excluded from variate analysis. The darker blue color indicates higher co-linearity.

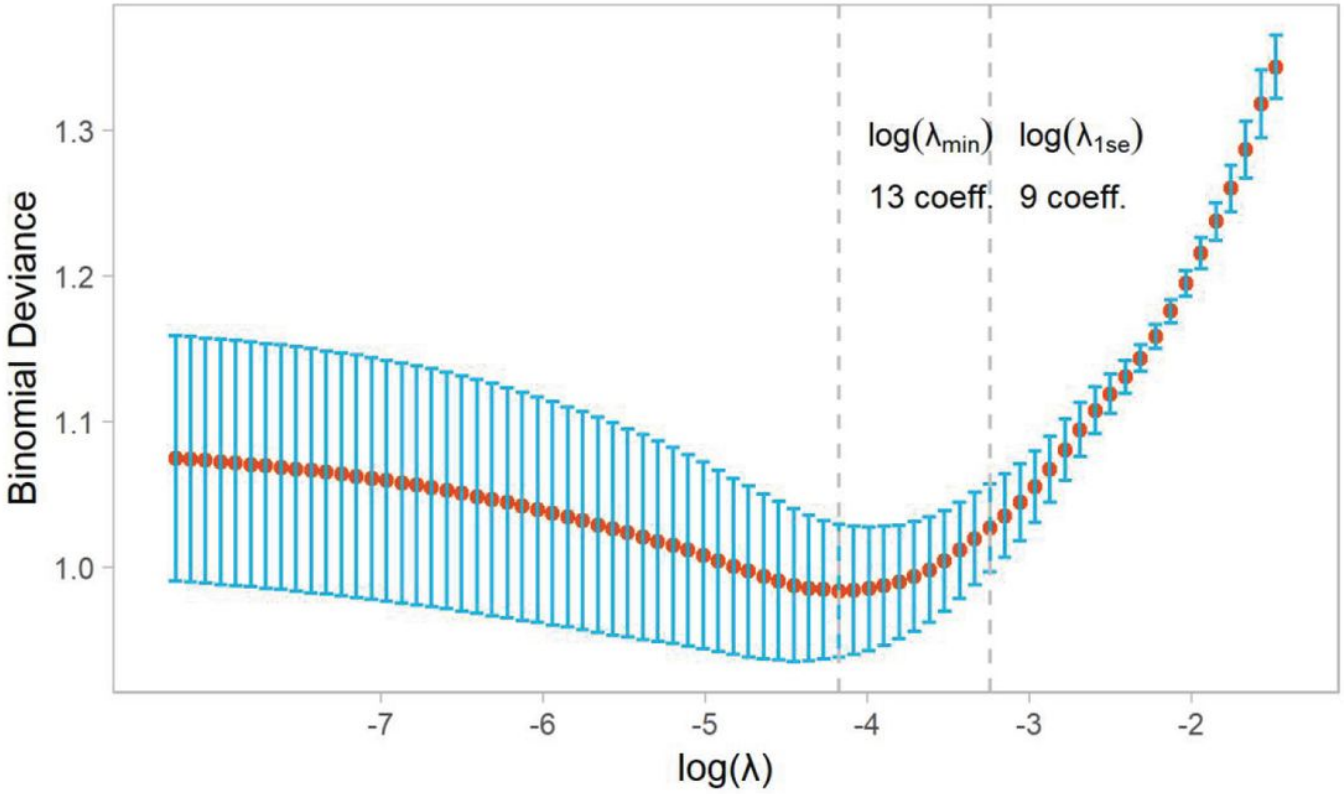

Figure 3

LASSO regression feature filtering

LASSO (Least Absolute Shrinkage and Selection Operator) regression based on five times ten-fold cross-validation was used for feature selection.

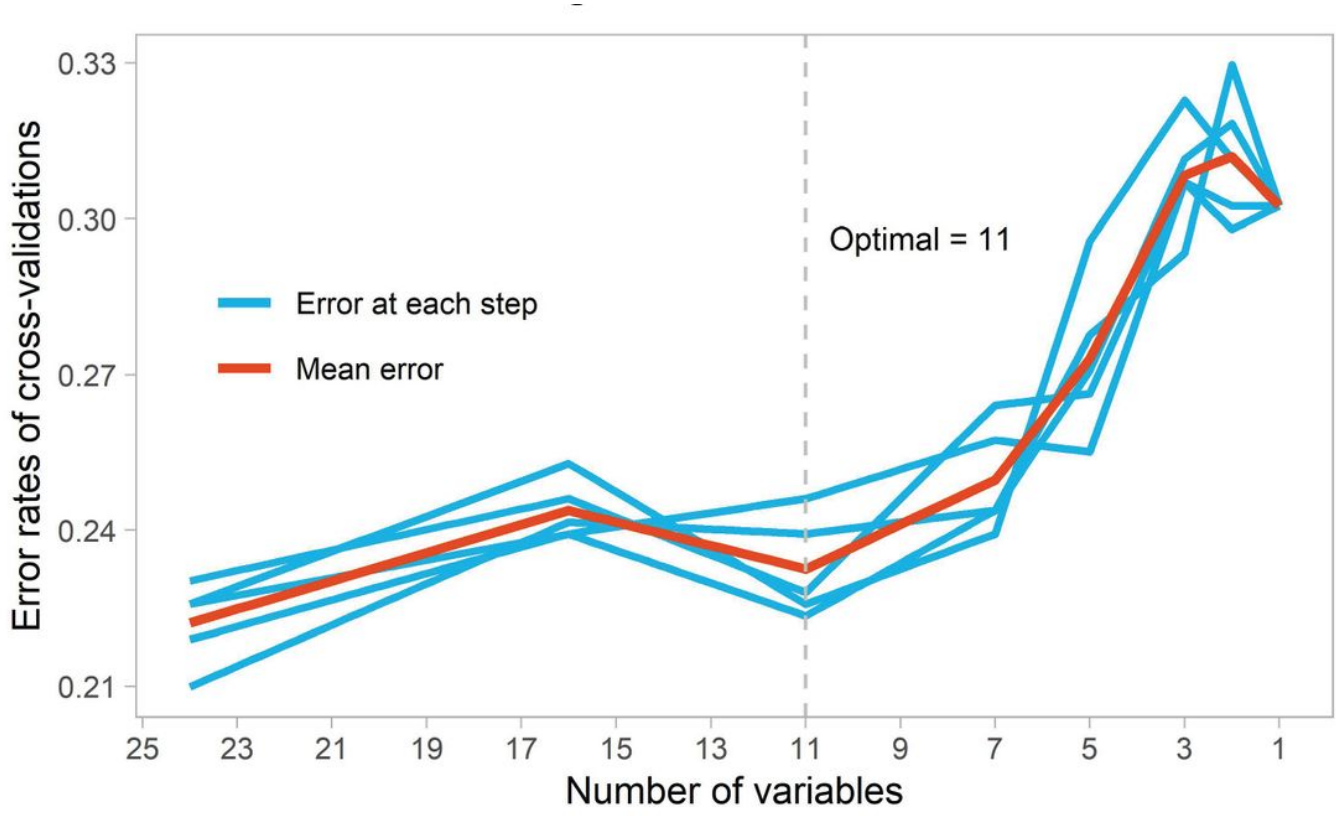

Figure 4

Random forest feature filtering

Random forest based on five times ten-fold cross-validation was used to perform feature selection.

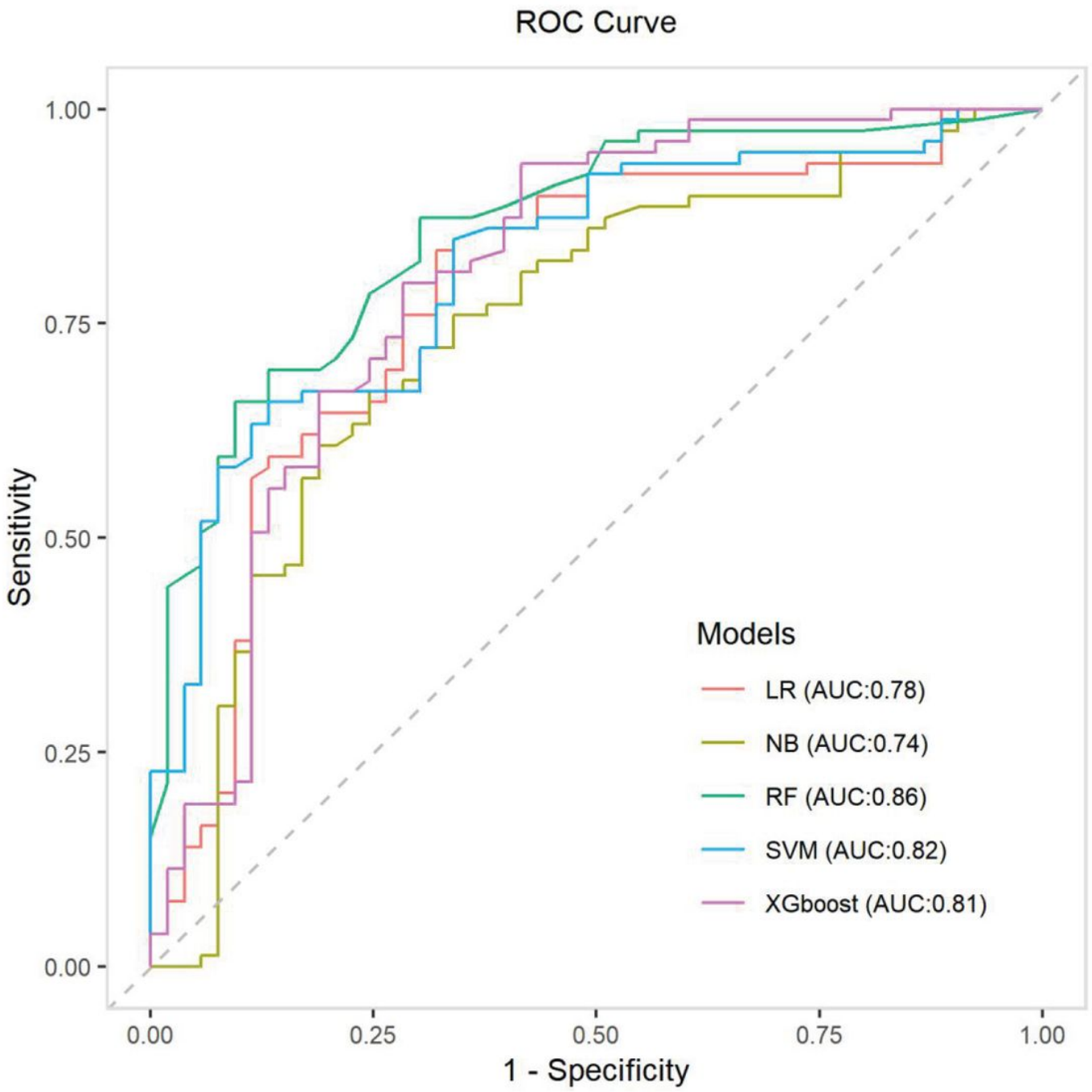

Figure 5

Receiver operating characteristic (ROC) curves of predictive models.

Diagnostic abilities of predictive models for the differential diagnosis of dMMR and pMMR in the test set. ROC curves of predictive model created by LR, NB, RF, SVM, and XGboost.

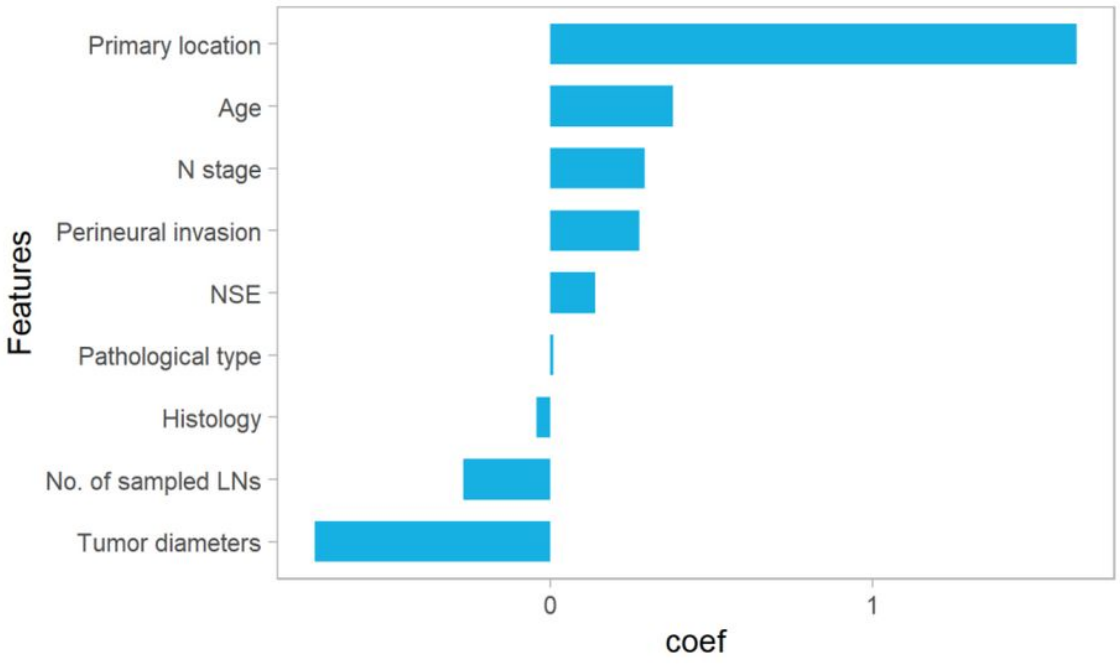

Figure 6

The initial variable importance analysis in Lasso Regression

For the LASSO (Least Absolute Shrinkage and Selection Operator) regression, we give the normalized regression coefficients for each feature.

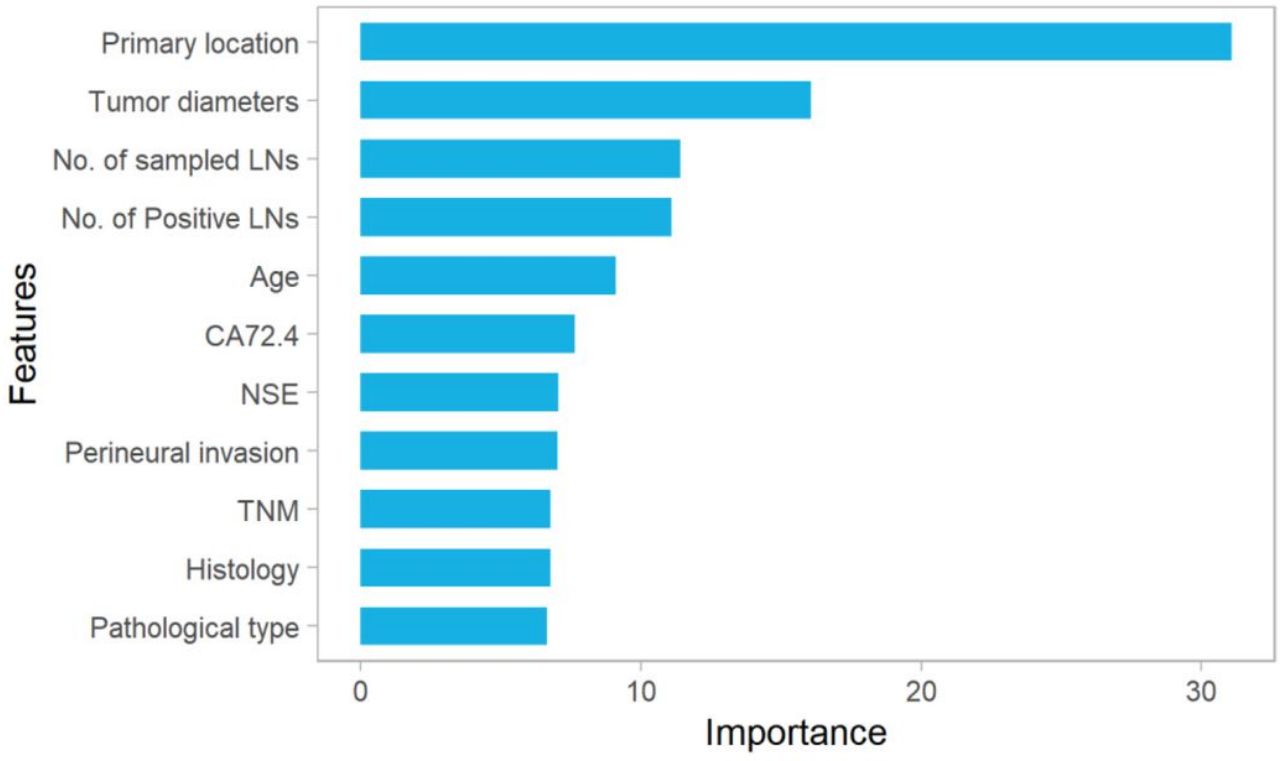

Figure 7

The initial variable importance analysis in Random forest

We used the machine learning technique, random forest, to determine feature importance.

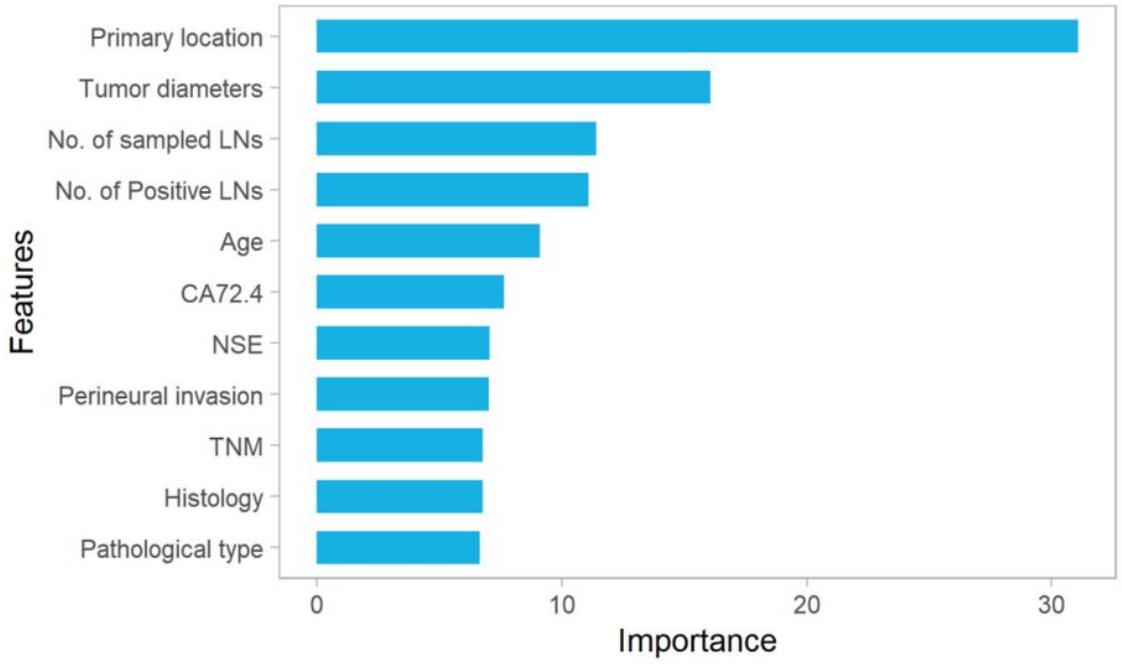

Figure 8

The final variable importance analysis in Random Forest

The merged variables were performed for feature importance analysis by random forest.

Supplementary Files

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

- SupplementaryFigure1.docx
- SupplementaryFigure2.docx
- SupplementaryFigure3.docx