Original Article

Construction of diagnostic models with machine-learning algorithms for colorectal cancer based on clinical laboratory parameters

Dengqing Si[1#], Yu Shu[1#], Hongbo Jiang[1], Xueping Lin[1#], Qiurong Yuan[1], Shaotuan Deng[1], Wei Luo[2], Yangze Lin[2], Ju Wang[2], Chengxiong Zhan[2], Aasma Shaukat[3], Peter C. Ambe[4,5], Shiqiong Niu[1], Zhaofan Luo[1]

[1]Department of Clinical Medical Laboratory, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, China; [2]Shenzhen Mindray Bio-Medical Electronics Co., Ltd., Shenzhen, China; [3]Division of Gastroenterology, NYU Grossman School of Medicine, New York, NY, USA; [4]Department of Surgery II, Witten/Herdecke University, Witten, Germany; [5]Department of General Surgery, Visceral Surgery and Coloproctology, GFO Kliniken Rhein Berg, Vinzenz-Pallotti-Hospital Bensberg, Bergisch Gladbach, Germany

Contributions: (I) Conception and design: D Si, Y Shu, X Lin; (II) Administrative support: Z Luo, S Niu; (III) Provision of study materials or patients: D Si, Y Shu, X Lin, H Jiang, Z Luo, S Niu; (IV) Collection and assembly of data: D Si, Y Shu, H Jiang, Q Yuan, S Deng, W Luo, Y Lin, J Wang, C Zhan; (V) Data analysis and interpretation: D Si, Y Shu, H Jiang, Q Yuan, S Deng, W Luo, Y Lin, J Wang, C Zhan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Zhaofan Luo, PhD; Shiqiong Niu, BS. Department of Clinical Medical Laboratory, The Seventh Affiliated Hospital of Sun Yat-sen University, No. 628 Zhenyuan Road, Shenzhen 518107, China. Email: luozhaofan@sysush.com; niushiqiong@sysush.com.

Background: Colonoscopy remains the predominant diagnostic modality for colorectal cancer (CRC), as the diagnostic performance of tumor markers in alone, particularly in the early stages of the disease, is limited. This study sought to develop a diagnostic model for CRC that integrated various laboratory parameters.

Methods: One hundred patients with CRC were assigned to an experimental group while 114 with benign colorectal diseases and 101 healthy individuals were assigned to a control group. The clinical and laboratory data, including the tumor markers such as carcinoembryonic antigen (CEA), glycan carbohydrate antigen 19-9 (CA19-9), carbohydrate antigen 242 (CA242), blood count parameters, blood biochemical parameters, and coagulation parameters, were collected for each participant. Three machine-learning models [multilayered perceptron (MLP), eXtreme Gradient Boosting (XGBoost), and random forest (RF)] were used to construct CRC diagnostic models. The performance of each model was evaluated based on its area under the curve (AUC), sensitivity, and specificity.

Results: There are 12 parameters: including CEA, CA19-9, CA242, absolute neutrophil value (NEUT), hemoglobin, the neutrophil/lymphocyte ratio, the platelet/lymphocyte ratio, alanine aminotransferase, alkaline phosphatase, aspartate aminotransferase, albumin, and prothrombin time, were selected to build the diagnostic model. For the validation set, the RF machine-learning model achieved the highest performance in identifying CRC [AUC: 0.902 (95% confidence interval: 0.812–0.989), accuracy: 0.803, sensitivity: 0.908, specificity: 0.772, positive predictive value: 0.664, negative predictive value: 0.890, and F1 score: 0.763]. The AUC, sensitivity, specificity, and Youden's index for the combined diagnosis of tumor markers CEA, CA19-9, and CA242 were 0.761, 0.486, 0.983, and 0.469, respectively. The RF diagnostic model showed better diagnostic efficacy than the combined diagnosis model of tumor markers CEA, CA19-9 and CA242.

Conclusions: The use of machine learning combined with multiple laboratory parameters effectively improved the diagnostic efficiency of CRC and provided more accurate results for clinical diagnosis.

Keywords: Colorectal cancer (CRC); machine learning; diagnostic model; tumor markers

Submitted Jul 05, 2024. Accepted for publication Aug 29, 2024. Published online Sep 12, 2024.
doi: 10.21037/jgo-24-516
View this article at: https://dx.doi.org/10.21037/jgo-24-516

Introduction

Colorectal cancer (CRC) is a common malignant tumor of the digestive tract. According to recent statistics, the worldwide incidence of CRC was 10% in 2020, second only to breast cancer in females and lung cancer, with a mortality rate of 9.4%, second only to lung cancer (1). CRC is also one of the most common malignant tumors in China. According to the China Cancer Statistics Report, CRC ranks second and fourth in terms of its incidence and mortality rates, respectively (2,3). In 2020, 555,000 new cases of CRC and 286,000 CRC-related deaths were reported (2,3). The incidence and mortality rates of CRC in China have shown an increasing trend in recent years (2,3).

Patients with CRC show no obvious symptoms at early stages, and half of CRC patients become symptomatic in the intermediate and advanced stages (4). Nearly 20–25% of patients have metastasis at the time of diagnosis (4). Even after radical surgery, one third of patients develop recurrent metastasis, thus the prognosis of CRC is poor (4). The five-year relative survival rate of patients with stage I CRC is 90%, and that of patients with stage IV CRC is only 14% (5).

It is believed that most CRCs originate from polyps. Polyps that are adenomatous acquire mutations and epigenetic changes over time, and eventually progress to CRC over a period of approximately 10 to 15 years (6). Previous studies have shown that regular CRC screening reduces the incidence and mortality of CRC (7,8). Therefore, the early screening, detection, diagnosis, and treatment of CRC are key to reducing the incidence and mortality of CRC (7,8).

Colonoscopy is the most sensitive test for detection and prevention of CRC. Adequate bowel prepping is required for optimal visualization of the mucosa during colonoscopy. However, the strenuous nature of bowel prepping has led to reduced patients' compliance, leading to low participation rates. Computed tomography colonography (CT colonography) is a non-invasive alternative to colonoscopy, especially for those unable to undergo colonoscopy. However, strict bowel preparation is also needed prior to CT colonography. More so, specialized equipment and personnel, and radiation risk make it unsuitable for screening large-scale populations. Nonetheless, CT colonography may not readily discriminate between residual faeces and true mucosal processes and smaller polyps are easily missed. More so, the therapeutic aspect of polypectomy during colonoscopy cannot be achieved with CT colonography (9). Fecal occult blood testing is commonly used in clinical practice. It is low-cost and non-invasive, but its sensitivity and detection rates are low (10). Serum tumor markers are produced by tumor cells or by the body in response to tumor cells. These markers reflect the existence and growth of tumor cells and play an important role in the diagnosis, follow-up, and recurrence monitoring of malignant tumors. One commonly used tumor marker is carcinoembryonic antigen (CEA), which plays an important role in the diagnosis and prognosis of CRC. Tumor markers are also elevated in benign diseases and can be at normal levels in malignant tumors. Thus, the limited specificity represents a major limitation for the use of tumor markers for diagnostic purposes (11).

In addition to the identification of novel tumor markers, diagnostic models can be developed that include

Highlight box

Key findings

- Using 12 parameters as inputs, diagnostic models were developed using the following three machine-learning algorithms: multilayered perceptron (MLP), random forest (RF), and eXtreme Gradient Boosting (XGBoost). The results showed that the diagnostic performance of these models was excellent, with the RF model demonstrating superior predictive capabilities compared to the other algorithms.

What is known, and what is new?

- Previous studies have indicated that when integrated with 24 parameters, the support vector machine algorithm demonstrates a strong diagnostic efficacy for colorectal cancer (CRC).
- Our study demonstrated that the RF algorithm achieved comparable diagnostic performance in the clinical diagnosis of CRC using only 12 parameters.

What is the implication, and what should change now?

- The RF algorithm demonstrated an ability to enhance the diagnostic effectiveness of CRC. Thus, this model could enable the more precise identification of individuals afflicted with this disease and increase the precision and sensitivity of early detection models.

a combination of already known markers to improve the diagnosis of CRC. Over the past decade, there has been tremendous progress in the technology of artificial intelligence. Clinical machine learning-based techniques can recognize patterns of high-dimensional data sets and help clinicians to make decisions for the early diagnosis and treatment of diseases (12,13). There is increasing evidence that the integration of pathological sections, blood markers, and machine-learning techniques can enhance the diagnostic accuracy of CRC models (14-16). With the extensive application of machine learning in the field of medical research, its potential and advantages for tumor prognosis and survival assessment have gradually advanced. Clinical laboratories can provide data resources for machine learning. Machine learning can extract rules or models related to biomarkers and clinical diseases and build complex and multi-parameter methods to assist in clinical decision making. In this study, biomarkers for CRC were analyzed by multilayered perceptron (MLP), eXtreme Gradient Boosting (XGBoost), and random forest (RF) algorithms to construct diagnostic models. We present this article in accordance with the TRIPOD reporting checklist (available at https://jgo.amegroups.com/article/view/10.21037/jgo-24-516/rc).

Methods

Study population

A retrospective study enrolled a total of 314 participants. A total of 100 CRC patients admitted to The Seventh Affiliated Hospital of Sun Yat-sen University from 2020 to 2022 were included in the study as the experimental group, of whom 60 were males and 40 were females. According to the tumor node metastasis (TNM) staging, 17 of these patients had stage I CRC, 27 had stage II CRC, 36 had stage III CRC, and 20 had stage IV CRC. Additionally, 214 healthy participants or participants with colorectal benign polyps were enrolled in the control group over the same period. Of the 214 participants in the control group, 113 had benign colorectal polyps, of whom 75 were males and 38 were females, aged 45–75 years, while the remaining 101 were healthy participants, of whom 70 were males and 31 were females. Both the CRC and colon polyps diagnoses were pathologically confirmed. All neoplasms, whether benign or malignant, were authenticated through pathological examination and immunodiagnostic procedures. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Ethics Committee of The Seventh Affiliated Hospital of Sun Yat-sen University (approval No. KY-2020-039-01, approval date: October 25, 2020). Individual informed consent was waived due to the retrospective nature of this study.

Inclusion criteria and exclusion criteria

The inclusion criteria for the study were as follows: (I) patients were included in the experimental group if they had been diagnosed with CRC after a pathological examination and had not been treated with surgery, radiotherapy, or chemotherapy; (II) patients were included in the control group if they had been pathologically diagnosed with benign colorectal polyps, and CRC had been ruled out. Patients were excluded from the study if they met any of the following exclusion criteria: (I) had a serious disease, such as heart, lung, liver, or kidney disease, severe immunodeficiency, serious infection, or any other disease; and/or (II) had a malignant tumor in another site, or had received radiotherapy or chemotherapy due to another malignant tumor.

Staging criteria

The clinical staging of CRC was performed based on the American Joint Committee on Cancer/Union Internationale Contre le Cancer (AJCC/UICC) TNM staging system for CRC (Eighth Edition, 2017) (17).

Laboratory data collection

According to the literature, the following laboratory data related to CRC were selected: (I) tumor markers, including CEA, carbohydrate antigen 19-9 (CA19-9), carbohydrate antigen 242 (CA242), and carbohydrate antigen 50 (CA50); (II) blood count parameters, including the white blood cell (WBC) count, absolute neutrophil value (NEUT), absolute lymphocyte value (LYM), absolute monocyte value (MON), platelet count (PLT), hemoglobin (HGB), and mean platelet volume (MPV); (III) blood biochemical parameters, including urea nitrogen (UREA), creatinine (CREA), uric acid (UA), total protein (TP), albumin (ALB), total bilirubin (TBIL), alanine aminotransferase (ALT), aspartate aminotransferase (AST), and alkaline phosphatase (ALP); and (IV) coagulation parameters, including prothrombin time (PT), activated partial thromboplastin time (APTT),

thrombin time (TT), fibrinogen (FIB), and the international normalized ratio (INR). Further, some indicators were converted into ratios, including the neutrophil/lymphocyte ratio (NLR) and platelet/lymphocyte ratio (PLR).

Data processing

The clinical laboratory data were randomly divided into two sets, and 70% (219 cases) of the data were used to train the model, and 30% (95 cases) of the data were used to verify the model.

Feature selection

The training set data were used for feature filtering. To improve the applicability of the model, and reduce the error caused by collinearity and correlation, Gaussian Naïve Bayes (GNB) classification, a neural network classification algorithm (i.e., MLP), and an Adaboost machine-learning algorithm were used to filter out the top 20 most important features for each of the models. Finally, the features were screened for model construction by machine-learning algorithms.

Machine model construction

We used a machine-learning technique to develop the following three models to diagnose CRC: MLP, RF, and XGBoost. We randomly divided the data set into the training and validation sets at a ratio of 7:3, cross-validated the training set five times, and verified the predictive ability of each model using the test set. The diagnostic performance of the three machine-learning models was evaluated based on the sensitivity, specificity, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve corresponding to the highest point of the Youden index.

Statistical analysis

SPSS 25.0 (IBM SPSS Statistics, Armonk, NK, USA) was used to analyze data. The normally distributed data were expressed as the mean ± standard deviation, and the *t*-test was used to compare the two groups. The non-normally distributed data were expressed as the median (interquartile range), and the Mann-Whitney *U* test was used to compare the two groups. Machine-learning models were performed using R version 3.6.3 and Python version 3.7. A P value <0.05 was considered statistically significant.

Results

Participant characteristics

A flowchart was used to delineate the subjects, clinical parameters, and methodological approaches utilized in the diagnosis of CRC (*Figure 1*). The baseline characteristics of the participants are presented in *Table 1*. There were no statistically significant differences between the experimental and control groups in terms of age and sex. The CEA level, CA242 level, NLR, and PLR were higher in the CRC patients than the control subjects, while HGB was significantly lower in the CRC patients than the control subjects.

Feature selection

GNB classification (*Figure 2A*), a neural network classification algorithm (i.e., MLP) (*Figure 2B*), and an Adaboost (*Figure 2C*) machine-learning algorithm were used to identify the top 20 feature variables of the three models. Ultimately, 12 parameters (i.e., CEA, CA19-9, CA242, NEUT, HGB, the NLR, the PLR, ALT, ALP, AST, ALB, and the PT) were identified to establish diagnostic models with machine-learning algorithms.

Model performance

After the feature screening, 12 parameters (i.e., CEA, CA19-9, CA242, NEUT, HGB, NLR, PLR, ALT, ALP, AST, ALB, and PT) were found to be significantly associated with the diagnosis of CRC. Based on these 12 parameters, three machine-learning algorithms (i.e., MLP, RF, and XGBoost) were used to construct the diagnostic models. The diagnostic performance of the three machine-learning models is shown in *Table 2*. The ROC curves were plotted to compare the diagnostic performance of the three machine-learning models (*Figure 3A*). The RF model had better prediction performance than the other machine-learning models (*Figure 3A*).

For the validation set (*Figure 3B*), the RF model had higher diagnostic efficacy than the MLP and XGBoost models. The accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and AUC [95% confidence interval (CI)] of the RF model, were 0.803, 0.908, 0.772, 0.664, 0.890, 0.763, and 0.902 (0.812–0.989), respectively.

Calibration plots with the Brier scores for the MLP, RF, and XGBoost models for CRC are shown in *Figure 4*. Notably, the

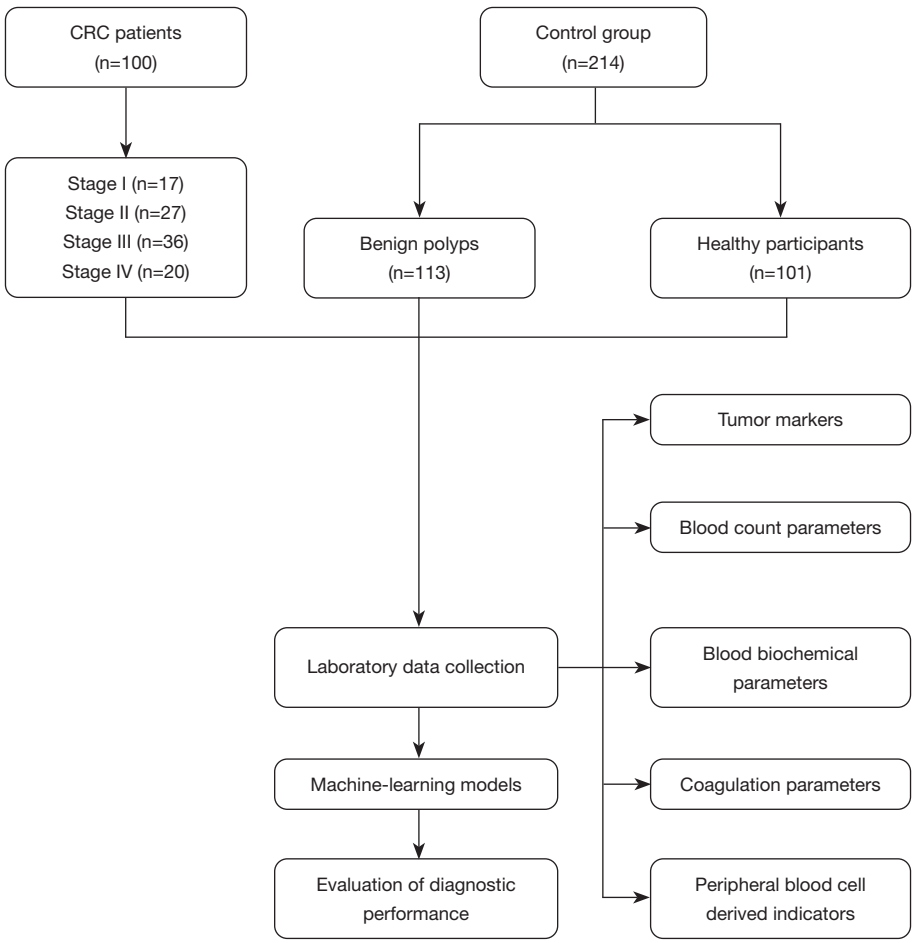

Figure 1 Flowchart of the subject recruitment and data collection. CRC, colorectal cancer.

RF model had the lowest Brier score for diagnosing CRC (0.097). The Brier scores for the MLP and XGBoost models were 0.100 and 0.112, respectively. The comparison of the performance of the machine-learning models, showed that the RF model had the best performance of all the models for diagnosing CRC.

The importance of the inspection indicators in the models

The top four indicators in the RF model were HGB, ALP, CEA, and ALB (*Figure 5*).

Comparison of the machine-learning models and tumor markers

The three machine-learning models were compared, and the results showed that the diagnostic performance of the RF model was better than that of the other two models. Thus, the diagnostic performance of the RF model was compared to CEA alone and to a combination of CEA, CA19-9, and CA242. In terms of the AUC, sensitivity, and specificity, the diagnosis efficacy of the RF model was higher than that of the traditional tumor markers. Thus, the models that combined machine-learning algorithms with several biomarkers significantly improved the diagnosis of CRC (*Table 3*).

Discussion

With high incidence and mortality rates, CRC is one of the most common malignant tumors in China (2,3). Due to changes in the living environment and eating habits of people in China (2,3), the incidence and mortality rate of CRC continue to increase. Thus, early screening is

Table 1 Participant characteristics

Baseline characteristic	Control group (n=214)	Experimental group (n=100)
Age (years)	55.50 (50.00, 62.50)	55.00 (52.00, 71.00)
Sex ratio (M/F)	144/70	60/40
CEA (ng/mL)	1.91 (1.25, 2.22)	2.20 (1.22, 10.51)*
CA19-9 (ng/mL)	10.79 (7.33, 13.78)	9.53 (5.58, 23.78)*
CA242 (ng/mL)	4.53 (2.91, 7.43)	5.25 (2.53, 23.00)*
CA50 (ng/mL)	5.98 (4.45, 8.10)	5.69 (3.09, 14.97)*
WBC ($\times10^9$/L)	6.82 (5.34, 7.95)	5.80 (4.70, 7.08)
NEUT WBC ($\times10^9$/L)	3.61 (2.98, 5.06)	3.22 (2.57, 4.14)
LYM WBC ($\times10^9$/L)	2.14 (1.82, 2.45)	1.86 (1.33, 2.28)*
MON WBC ($\times10^9$/L)	0.42 (0.32, 0.52)	0.37 (0.32, 0.52)
PLT WBC ($\times10^9$/L)	237.0 (221.0, 277.5)	247.0 (199.0, 311.0)
HGB (g/L)	143.0 (124.5, 157.0)	122.0 (104.0, 135.0)*
MPV (fL)	9.80 (8.93, 10.00)	10.00 (8.95, 11.00)
NLR	1.63 (1.24, 2.76)	1.78 (1.33, 2.31)*
PLR	114.08 (95.47, 123.26)	145.45 (116.67, 189.04)*
UREA (mmol/L)	4.80 (3.93, 5.53)	4.60 (3.60, 5.80)
CREA (μmol/L)	79.50 (65.50, 83.25)	66.00 (59.67, 81.00)
UA (mmol/L)	333.75 (280.68, 382.58)	310.60 (272.80, 385.80)*
TP (g/L)	68.70 (66.33, 73.05)	68.20 (62.80, 71.00)*
ALB (g/L)	40.45 (39.10, 43.18)	39.40 (38.00, 40.60)*
TBIL (μmol/L)	13.95 (8.83, 20.72)	9.09 (6.84, 15.80)*
ALT (U/L)	16.50 (12.75, 26.00)	14.00 (10.00, 23.00)
AST (U/L)	20.50 (16.00, 25.50)	17.00 (14.00, 20.00)
ALP (U/L)	69.50 (58.75, 84.25)	66.00 (56.00, 77.00)
PT (s)	11.65 (11.20, 12.30)	11.80 (11.40, 12.50)*
APTT (s)	26.85 (25.30, 28.80)	27.40 (25.90, 28.60)
TT (s)	17.70 (16.98, 18.10)	17.30 (16.80, 17.90)*
FIB (g/L)	2.60 (2.33, 3.09)	2.69 (2.40, 3.44)*
INR	0.97 (0.94, 1.03)	0.99 (0.96, 1.05)*

Data are presented as numbers or median (Q1, Q3). *, P<0.05, versus the control group. CEA, carcinoembryonic antigen; CA19-9, carbohydrate antigen 19-9; CA242, carbohydrate antigen 242; CA50, carbohydrate antigen 50; WBC, white blood cell count; NEUT, absolute neutrophil value; LYM, absolute lymphocyte value; MON, absolute monocyte value; PLT, platelet count; HGB, hemoglobin; MPV, mean platelet volume; NLR, neutrophil/lymphocyte ratio; PLR, platelet/lymphocyte ratio; UREA, urea nitrogen; CREA, creatinine; UA, uric acid; TP, total protein; ALB, albumin; TBIL, total bilirubin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALP, alkaline phosphatase; PT, prothrombin time; APTT, activated partial thromboplastin time; TT, thrombin time; FIB, fibrinogen; INR, international normalized ratio; Q1, 25th percentile; Q3, 75th percentile.

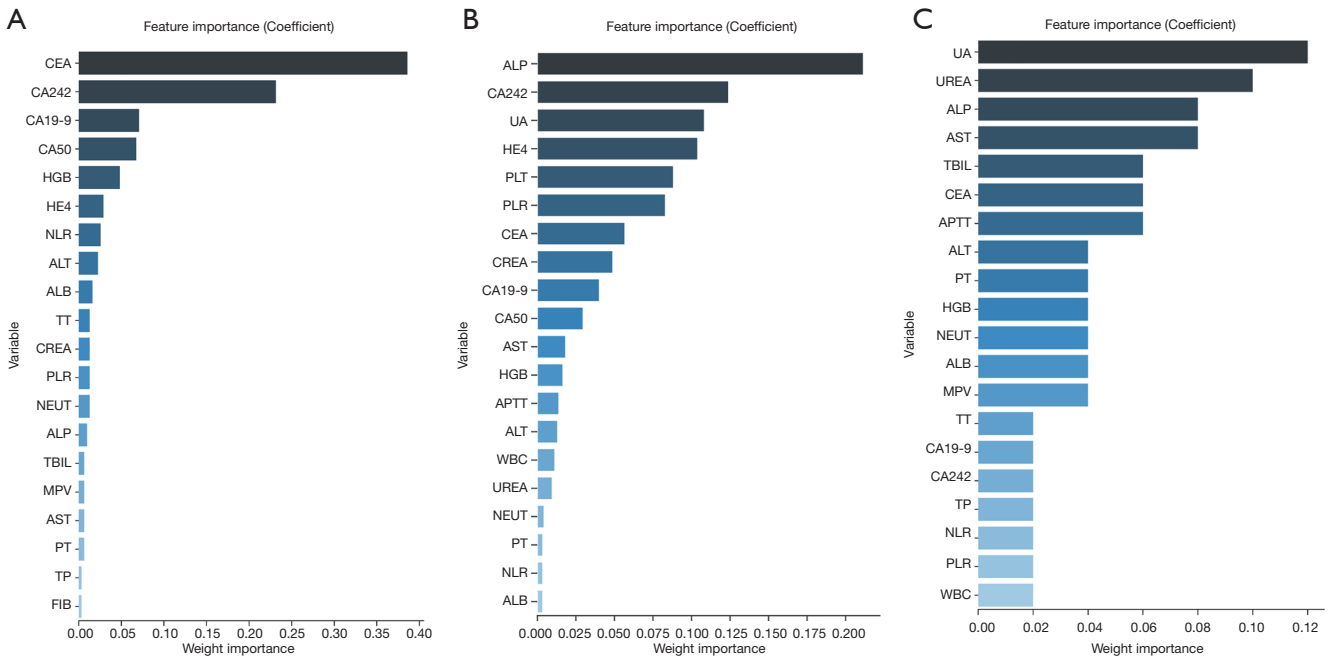

Figure 2 Importance ranking graphs showing the impact factors for the following three machine-learning models: (A) GNB; (B) MLP; and (C) Adaboost. CEA, carcinoembryonic antigen; CA242, carbohydrate antigen 242; CA19-9, carbohydrate antigen 19-9; CA50, carbohydrate antigen 50; HGB, hemoglobin; HE4, human epididymis protein 4; NLR, neutrophil/lymphocyte ratio; ALT, alanine aminotransferase; ALB, albumin; TT, thrombin time; CREA, creatinine; PLR, platelet/lymphocyte ratio; NEUT, absolute neutrophil value; ALP, alkaline phosphatase; TBIL, total bilirubin; MPV, mean platelet volume; AST, aspartate aminotransferase; PT, prothrombin time; TP, total protein; FIB, fibrinogen; UA, uric acid; PLT, platelet count; APTT, activated partial thromboplastin time; WBC, white blood cell count; UREA, urea nitrogen; GNB, Gaussian Plain Bayesian; MLP, multilayered perceptron.

Table 2 Diagnostic efficacy of the three machine-learning models for the training and validation sets

Data set	Classification model	AUC	Accuracy	Sensitivity	Specificity	Positive predictive value	Negative predictive value	F1 score
Training set	RF	0.974 (0.955–0.993)	0.904 (0.873–0.936)	0.904 (0.848–0.960)	0.912 (0.846–0.979)	0.832 (0.720–0.944)	0.951 (0.931–0.971)	0.858 (0.822–0.894)
	XGBoost	0.951 (0.921–0.981)	0.897 (0.888–0.907)	0.827 (0.793–0.861)	0.935 (0.911–0.959)	0.844 (0.798–0.891)	0.921 (0.910–0.931)	0.834 (0.822–0.845)
	MLP	0.899 (0.854–0.944)	0.817 (0.794–0.841)	0.838 (0.758–0.919)	0.817 (0.768–0.866)	0.658 (0.618–0.699)	0.917 (0.886–0.948)	0.734 (0.701–0.767)
Validation set	RF	0.902 (0.812–0.989)	0.803 (0.729–0.877)	0.908 (0.851–0.964)	0.772 (0.710–0.834)	0.664 (0.543–0.785)	0.890 (0.838–0.943)	0.763 (0.666–0.860)
	XGBoost	0.891 (0.792–0.989)	0.808 (0.747–0.869)	0.815 (0.713–0.918)	0.849 (0.723–0.975)	0.725 (0.566–0.884)	0.857 (0.792–0.921)	0.754 (0.670–0.838)
	MLP	0.838 (0.711–0.965)	0.758 (0.685–0.831)	0.800 (0.645–0.955)	0.837 (0.691–0.983)	0.572 (0.477–0.667)	0.874 (0.816–0.933)	0.654 (0.568–0.740)

Data are presented as numerical value with 95% CI. RF, random forest; XGBoost, eXtreme Gradient Boosting; MLP, multilayered perceptron; AUC, area under the curve; CI, confidence interval.

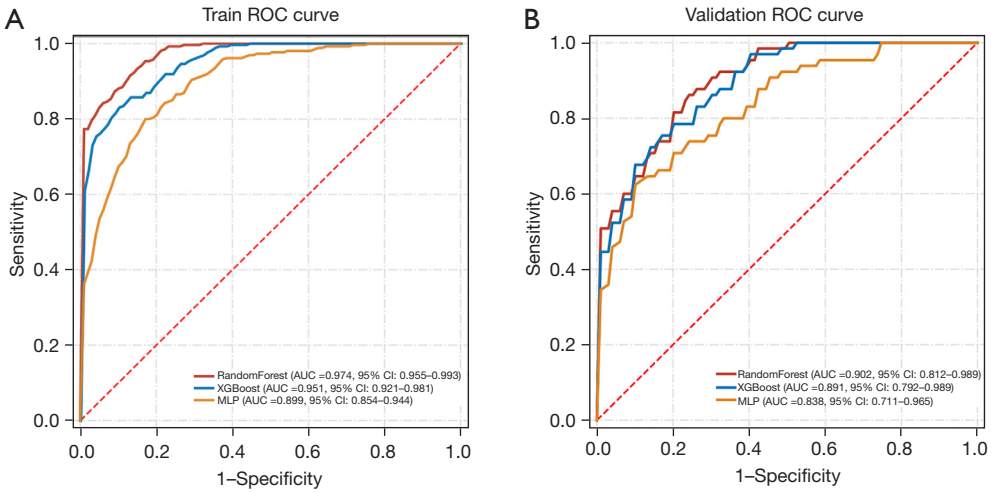

Figure 3 ROC curves for the training and validation sets of the three machine-learning models. ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval; XGBoost, eXtreme Gradient Boosting; MLP, multilayered perceptron.

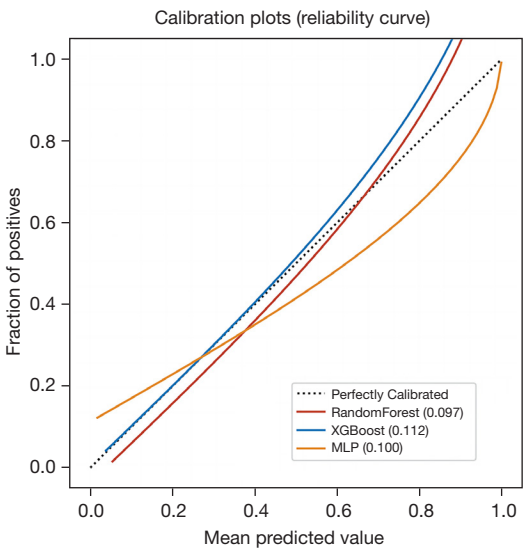

Figure 4 Validation set calibration curves of the three machine-learning models. XGBoost, eXtreme Gradient Boosting; MLP, multilayered perceptron.

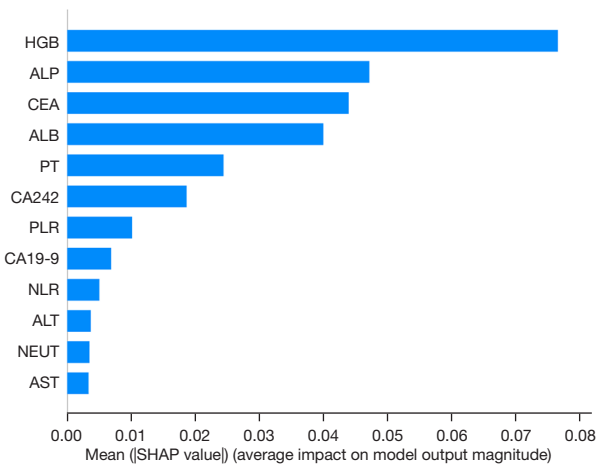

Figure 5 RF model SHAP importance graph. HGB, hemoglobin; ALP, alkaline phosphatase; CEA, carcinoembryonic antigen; ALB, albumin; PT, prothrombin time; CA242, carbohydrate antigen 242; PLR, platelet/lymphocyte ratio; CA19-9, carbohydrate antigen 19-9; NLR, neutrophil/lymphocyte ratio; ALT, alanine aminotransferase; NEUT, absolute neutrophil value; AST, aspartate aminotransferase; SHAP, shapely additive explanations; RF, random forest.

important for the treatment and prognosis of CRC patients. Of the three machine-learning models investigated, the RF model outperformed the other two models with regard to diagnostic accuracy, sensitivity, specificity, positive and negative prediction values, F1 score and an AUC of 0.902.

In this study, we utilized routine laboratory data to develop machine learning models aimed at identifying CRC patients, employing MLP, XGBoost, and RF for model construction. MLP, also known as an artificial neural network (ANN), uses the basic principles of neural networks in biology. ANN examines network topology as the theoretical foundation for simulating the processing of complex information within the human nervous system. It uses a multi-level model

Table 3 Diagnostic efficacy of the RF machine-learning model and tumor markers

Variables	AUC	Sensitivity	Specificity	Youden index
CEA	0.740	0.629	0.783	0.412
CEA + CA19-9 + CA242	0.761	0.486	0.983	0.469
RF model	0.864	0.829	0.800	0.629

Data are presented as numerical value. CEA, carcinoembryonic antigen; CA19-9, carbohydrate antigen 19-9; CA242, carbohydrate antigen 242; RF, random forest; AUC, area under the curve.

composed of multiple neurons of the perceptron. It has parallel distributed processing power, high fault tolerance, and intelligence, and is capable of self-learning.

Baxter *et al.* (18) used four machine-learning models (i.e., XGboost, ANN, support vector machine, and RF) to mine the routine test data of enrolled patients. The AUCs of the XGboost model in the training set and the prospective validation set were 0.799 and 0.816, respectively, which were better than those of the fecal occult blood test. In another study, Li *et al.* (19) reported that the diagnostic AUC of a logistic regression model based on CEA, HGB, lipoprotein(a), and high-density lipoprotein was 0.849. In the present study, compared with the MLP and XGboost models, the RF model had better diagnostic performance, and its accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and AUC (95% CI) were 0.803, 0.908, 0.772, 0.664, 0.890, 0.763, and 0.902 (0.812–0.989), respectively. These figures suggest that RF may be a better predictive model in comparison to XGboost and MLP models.

A major strength of this study is the use of readily available data from routine clinical work-up. Twelve routinely investigated parameters i.e., CEA, CA19-9, CA242, NEUT, HGB, NLR, PLR, ALT, ALP, AST, ALB, and PT for CRC were identified and used in this study. CRC patients have diverse clinical manifestations, of which anemia is one of the most common clinical symptoms. HGB, which is an important indicator of anemia, is an independent predictor of CRC and can be used to screen CRC. HGB has certain clinical value in diagnosing incipient CRC (20,21).

Also, routinely measured tumor markers like CEA, which is usually elevated in, nearly 43–69% of CRC, as well as CA19-9 and CA 242 were interpreted in the models investigated in our study. Although different organs have different expression levels of CA19-9 (22). Its levels have been shown to correlate with tumor size, the depth of invasion, and lymph node metastasis. Thus, CA 19-9 is an auxiliary indicator of CRC (17-19,22-24). Similarly, CA242, although commonly used to diagnose pancreatic cancer, is expressed by CRC in a stage dependent fashion (25).

The nation of combining the three tumor markers is backed by data from Rao *et al.* (26), who reported CEA, CA242, and CA199 as important predictors of CRC risk. The inflammatory environment provides a favorable condition for the tumor growth, invasion, and metastasis of CRC (27). Inflammation can affect each stage of CRC development and regulate the polarization of tumor cells (28,29). The NEUT, PLR, and NLR reflect the severity of inflammation, and studies have shown that neutrophils, the PLR, and the NLR are related to the prognosis and survival of CRC. High levels of inflammatory mediators predict the progression of CRC (30,31). Thus, incorporating these cellular markers in this study was in accordance with current clinical practice.

ALB is an acute phase reaction protein whose level decreases in the inflammatory response. Low ALB levels indicate tumor-induced malnutrition and prognosticate adverse outcomes (32,33). In addition, the liver metastasis of CRC accounts for about 70% of CRC metastasis, and the levels of ALB, ALT, AST, and ALP can reflect the degree of liver dysfunction and exhibit an inverse correlation with the risk of CRC (34).

Malignant tumors can increase the risk of venous thrombosis. Malignant tumor cells and their products interact with host cells to induce a hypercoagulable state, leading to thrombosis. Patients with cancer have a four-to-six-fold increased risk of thrombosis compared with patients without cancer, which reduces survival (35-37). The 12 indicators included in this study accounted for the occurrence and development of CRC, tumor inflammatory environment, thrombosis, etc., and had a diagnostic efficacy higher than that of traditional tumor markers such as CEA, CA19-9, and CA242. These findings are largely in line with the clinical experience with regard to the distribution of CRC in the general population.

A major limitation of this study is the small size of the experimental population. It is questionable if the results recorded in this study may be reproduced in a much larger population. Also, a stratification of the results generated based on the AJCC/UICC staging system was not performed, due to the small size of the respective stage groups. Despite these limitations, the potential implication of such models in clinical practice should encourage further investigations in the field. As previously mentioned, RF does not demonstrate a definitive superiority over XGBoost in terms of diagnostic performance for CRC, particularly in validation datasets. Nonetheless, when considering absolute metrics, RF outperforms XGBoost in diagnostic accuracy. Additionally, both RF and XGBoost algorithms have achieved high levels of accuracy in diagnosing CRC, indicating their advantages in this application. The significance of machine learning algorithms in the context of CRC diagnosis is underscored, and these findings represent advancements in clinical practice. The results suggest that the RF algorithm, in particular, may offer promising prospects for further research and development in this area.

Conclusions

In summary, this study used simple routine laboratory data to construct a diagnostic model for CRC. We found that the RF model had high sensitivity and specificity in diagnosing CRC. Thus, it could serve as a non-invasive and efficient method for identifying CRC.

Acknowledgments

We would like to thank Professor Zhao-Fan Luo for his writing assistance and proof reading the article.
Funding: None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://jgo.amegroups.com/article/view/10.21037/jgo-24-516/rc

Data Sharing Statement: Available at https://jgo.amegroups.com/article/view/10.21037/jgo-24-516/dss

Peer Review File: Available at https://jgo.amegroups.com/article/view/10.21037/jgo-24-516/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jgo.amegroups.com/article/view/10.21037/jgo-24-516/coif). A.S. reports consulting fees from Freenome Inc. and Iterative Health. W.L., Y.L., J.W., C.Z. are from Shenzhen Mindray Bio-Medical Electronics Co., Ltd., Shenzhen, China. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was approved by the Ethics Committee of The Seventh Affiliated Hospital of Sun Yat-sen University (approval No. KY-2020-039-01, approval date: October 25, 2020). Individual informed consent was waived due to the retrospective nature of this study. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
2. Zheng RS, Zhang SW, Sun KX, et al. Cancer statistics in China, 2016. Zhonghua Zhong Liu Za Zhi 2023;45:212-20.
3. Expert consensus on the early diagnosis and treatment of colorectal cancer in China (2023 edition). Zhonghua Yi Xue Za Zhi 2023;103:3896-908.
4. Joo HJ, Seok JU, Kim BC, et al. Effects of prior endoscopic resection on recurrence in patients with T1 colorectal cancer who underwent radical surgery. Int J Colorectal Dis 2023;38:167.
5. China guideline for the screening, early detection and early treatment of colorectal cancer (2020, Beijing). Zhonghua Zhong Liu Za Zhi 2021;43:16-38.
6. van Toledo DEFWM, IJspeert JEG, Spaander MCW, et

al. Colorectal cancer risk after removal of polyps in fecal immunochemical test based screening. EClinicalMedicine 2023;61:102066.
7. Barnell EK, Wurtzler EM, La Rocca J, et al. Multitarget Stool RNA Test for Colorectal Cancer Screening. JAMA 2023;330:1760-8.
8. Murphy CC, Zaki TA. Changing epidemiology of colorectal cancer - birth cohort effects and emerging risk factors. Nat Rev Gastroenterol Hepatol 2024;21:25-34.
9. Sosna J, Morrin MM, Kruskal JB, et al. CT colonography of colorectal polyps: a metaanalysis. AJR Am J Roentgenol 2003;181:1593-8.
10. Blom J, Saraste D, Törnberg S, et al. Routine Fecal Occult Blood Screening and Colorectal Cancer Mortality in Sweden. JAMA Netw Open 2024;7:e240516.
11. Bu F, Cao S, Deng X, et al. Evaluation of C-reactive protein and fibrinogen in comparison to CEA and CA72-4 as diagnostic biomarkers for colorectal cancer. Heliyon 2023;9:e16092.
12. Kourou K, Exarchos KP, Papaloukas C, et al. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Comput Struct Biotechnol J 2021;19:5546-55.
13. Wang Q, Xu J, Wang A, et al. Systematic review of machine learning-based radiomics approach for predicting microsatellite instability status in colorectal cancer. Radiol Med 2023;128:136-48.
14. Neto PC, Montezuma D, Oliveira SP, et al. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. NPJ Precis Oncol 2024;8:56.
15. Tsai PC, Lee TH, Kuo KC, et al. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. Nat Commun 2023;14:2102.
16. Krishnan ST, Winkler D, Creek D, et al. Staging of colorectal cancer using lipid biomarkers and machine learning. Metabolomics 2023;19:84.
17. Hari DM, Leung AM, Lee JH, et al. AJCC Cancer Staging Manual 7th edition criteria for colon cancer: do the complex modifications improve prognostic assessment? J Am Coll Surg 2013;217:181-90.
18. Baxter SL, Saseendrakumar BR, Paul P, et al. Predictive Analytics for Glaucoma Using Data From the All of Us Research Program. Am J Ophthalmol 2021;227:74-86.
19. Li H, Lin J, Xiao Y, et al. Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data. Technol Cancer Res Treat 2021;20:15330338211058352.
20. Kortlever T, de Klaver W, van der Vlugt M, et al. Cross-sectional risk models using quantitative fecal hemoglobin in colorectal cancer screening: a systematic review. Expert Rev Mol Diagn 2023;23:1221-32.
21. Krishnamoorthy A, Arasaradnam R. Colorectal cancer diagnostic biomarkers: Beyond faecal haemoglobin. Best Pract Res Clin Gastroenterol 2023;66:101870.
22. Hou S, Jing J, Wang Y, et al. Evaluation of Clinical Diagnostic and Prognostic Value of Preoperative Serum Carcinoembryonic Antigen, CA19-9, and CA24-2 for Colorectal Cancer. Altern Ther Health Med 2023;29:192-7.
23. Tang Y, Cui Y, Zhang S, et al. The sensitivity and specificity of serum glycan-based biomarkers for cancer detection. Prog Mol Biol Transl Sci 2019;162:121-40.
24. He Y, He X, Zhou Y, et al. Clinical value of circulating tumor cells and hematological parameters in 617 Chinese patients with colorectal cancer: retrospective analysis. BMC Cancer 2023;23:707.
25. Luo H, Shen K, Li B, et al. Clinical significance and diagnostic value of serum NSE, CEA, CA19-9, CA125 and CA242 levels in colorectal cancer. Oncol Lett 2020;20:742-50.
26. Rao H, Wu H, Huang Q, et al. Clinical Value of Serum CEA, CA24-2 and CA19-9 in Patients with Colorectal Cancer. Clin Lab 2021.
27. Tuomisto AE, Mäkinen MJ, Väyrynen JP. Systemic inflammation in colorectal cancer: Underlying factors, effects, and prognostic significance. World J Gastroenterol 2019;25:4383-404.
28. Schmitt M, Greten FR. The inflammatory pathogenesis of colorectal cancer. Nat Rev Immunol 2021;21:653-67.
29. Nadeem MS, Kumar V, Al-Abbasi FA, et al. Risk of colorectal cancer in inflammatory bowel diseases. Semin Cancer Biol 2020;64:51-60.
30. Zhang L, Shi FY, Qin Q, et al. Relationship between preoperative inflammatory indexes and prognosis of patients with rectal cancer and establishment of prognostic nomogram prediction model. Zhonghua Zhong Liu Za Zhi 2022;44:402-9.
31. Ergen ŞA, Barlas C, Yıldırım C, et al. Prognostic Role of Peripheral Neutrophil-Lymphocyte Ratio (NLR) and Platelet-Lymphocyte Ratio (PLR) in Patients with Rectal Cancer Undergoing Neoadjuvant Chemoradiotherapy. J Gastrointest Cancer 2022;53:151-60.
32. Sekiguchi K, Matsuda A, Yamada M, et al. The utility of serum osteopontin levels for predicting postoperative complications after colorectal cancer surgery. Int J Clin

Oncol 2022;27:1706-16.
33. Yamamoto T, Kawada K, Obama K. Inflammation-Related Biomarkers for the Prediction of Prognosis in Colorectal Cancer Patients. Int J Mol Sci 2021;22:8002.
34. He MM, Fang Z, Hang D, et al. Circulating liver function markers and colorectal cancer risk: A prospective cohort study in the UK Biobank. Int J Cancer 2021;148:1867-78.
35. Gulati S, Eckman MH. Anticoagulant Therapy for Cancer-Associated Thrombosis : A Cost-Effectiveness Analysis. Ann Intern Med 2023;176:1-9.
36. Wang TF, Carrier M, Carney BJ, et al. Anticoagulation management and related outcomes in patients with cancer-associated thrombosis and thrombocytopenia: A systematic review and meta-analysis. Thromb Res 2023;227:8-16.
37. Xu Y, Cole K, Collins E, et al. Anticoagulation for the Prevention of Arterial Thrombosis in Ambulatory Cancer Patients: Systematic Review and Meta-Analysis. JACC CardioOncol 2023;5:520-32.

Cite this article as: Si D, Shu Y, Jiang H, Lin X, Yuan Q, Deng S, Luo W, Lin Y, Wang J, Zhan C, Shaukat A, Ambe PC, Niu S, Luo Z. Construction of diagnostic models with machine-learning algorithms for colorectal cancer based on clinical laboratory parameters. J Gastrointest Oncol 2024;15(5):2145-2156. doi: 10.21037/jgo-24-516