Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data

Hui Li, MS[1,#], Jianmei Lin, BS[1,#], Yanhong Xiao, MS[1],
Wenwen Zheng, MS[1], Lu Zhao, PhD[1], Xiangling Yang, PhD[1,2],
Minsheng Zhong, MS[3], and Huanliang Liu, MD, PhD[1,2] (ID)

Abstract

Background: Current diagnostic methods for colorectal cancer (CRC) are colonoscopy and sigmoidoscopy, which are invasive and complex procedures with possible complications. This study aimed to determine models for CRC identification that involve minimally invasive, affordable, portable, and accurate screening variables. **Methods:** This was a retrospective study that used data from electronic medical records of patients with CRC and healthy individuals between July 2017 and June 2018. Laboratory data, including liver enzymes, lipid profiles, complete blood counts, and tumor biomarkers, were extracted from the electronic medical records. Five machine learning models (logistic regression, random forest, k-nearest neighbors, support vector machine [SVM], and naïve Bayes) were used to identify CRC. The performances were evaluated using the areas under the curve (AUCs), sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV). **Results:** A total of 1164 electronic medical records (CRC patients: 582; healthy controls: 582) were included. The logistic regression model achieved the highest performance in identifying CRC (AUC: 0.865, sensitivity: 89.5%, specificity: 83.5%, PPV: 84.4%, NPV: 88.9%). The first four weighted features in the model were carcinoembryonic antigen (CEA), hemoglobin (HGB), lipoprotein (a) (Lp(a)), and high-density lipoprotein (HDL). A diagnostic model for CRC was established based on the four indicators, with an AUC of 0.849 (0.840-0.860) for identifying all CRC patients, and it performed best in discriminating patients with late colon cancer from healthy individuals with an AUC of 0.905 (0.889-0.929). **Conclusions:** The logistic regression model based on CEA, HGB, Lp(a), and HDL might be a powerful, noninvasive, and cost-effective method to identify CRC.

Keywords

diagnosis, colorectal cancer, machine learning, logistic regression, clinical laboratory techniques

Abbreviations

AFP, α-fetoprotein; ALT, alanine transaminase; Apo, apolipoprotein; AST, aspartate transaminase; AUC, area under the curve; BMI, body mass index; CEA, carcinoembryonic antigen; TC, total cholesterol; CI, confidence interval; CRC, colorectal cancer; EMR, electronic medical record; ESO, eosinophils; gFOBT, guaiac fecal occult blood test; GGT, γ-glutamyl transferase; HDL,

[1] Department of Clinical Laboratory, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong, China
[2] Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong, China
[3] Department of Artificial Intelligence Laboratory, Xuanwu Technology, Guangzhou, Guangdong, China

[#] These authors contributed equally to this work.

Corresponding Authors:
Huanliang Liu, Department of Clinical Laboratory, Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Guangdong Institute of Gastroenterology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong 510655, China.
Email: liuhuanl@mail.sysu.edu.cn
Minsheng Zhong, Department of Artificial Intelligence Laboratory, Xuanwu Technology, Guangzhou, Guangdong 510620, China.
Email: bihai98@163.com

high-density lipoprotein; HGB, hemoglobin; hs-CRP, high-sensitivity C-reactive protein; LDL, low-density lipoprotein; Lp(a), lipoprotein (a); LYM, lymphocytes; MONO, monocytes; NPV, negative predictive value; NEU, neutrophils; PPV, positive predictive value; PLT, platelets; RBC, red blood cells; TG, triglycerides; WBC, white blood cells.

Introduction

Colorectal cancer (CRC) is the third most common cancer in men and the second most common in women.[1] Most CRCs occur sporadically, and the carcinogenesis process involves precancerous lesions, adenoma, carcinoma, and metastasis, accompanied by a series of abnormal metabolisms.[2] These long-term pathological and metabolic variations can allow the diagnosis of CRC.[3]

The current first-line detection methods for CRC are colonoscopy and sigmoidoscopy.[4] Still, they are invasive, complex procedures involving bowel preparation, along with possible severe complications such as hemorrhage, colonic perforation, and cardiorespiratory problems.[5] Metabolic variations in the blood can bring us some clues for the detection of CRC. The carcinoembryonic antigen (CEA) has been identified as a biomarker for CRC. It is the most frequent indicator of CRC recurrence in asymptomatic patients and currently is the most cost-effective test for the preclinical detection of resectable CRC, with a sensitivity of about 80% and a specificity of about 70%, and can provide a lead time of approximately 5 months.[6] Nevertheless, CEA alone is of little use in detecting early CRC, and high preoperative concentrations of CEA correlate with an adverse prognosis.[7] Lower hemoglobin (HGB) and red blood cells (RBC) levels, probably due to occult gastrointestinal hemorrhage in CRC, are associated with CRC.[8] van Duijnhoven et al.[9] revealed that the concentrations of high-density lipoproteins (HDL) were inversely associated with the risk of CRC. Furthermore, dyslipidemia, including hypertriglyceridemia and high low-density lipoproteins (LDL), has been associated with the development of CRC,[10,11] while the atherogenic lipoprotein (a) (Lp(a)) appears to be protective against CRC development.[12] Since liver metastasis accounts for 70% of CRC metastases, α-fetoprotein (AFP) levels may be elevated in CRC patients, and the alanine transaminase (ALT), aspartate transaminase (AST), and γ-glutaryl transferase (GGT) levels might be associated with severely impaired liver function.[13] Elevated platelets (PLT) are associated with the presence of CRC[14] and with CRC metastasis.[15] Inflammation and a chronic inflammatory state are associated with cancer development,[16] which can be reflected by C-reactive protein (CRP) levels.[17] Regarding the immune cells, elevated neutrophils (NEU) and lower lymphocytes (LYM) are associated with CRC and its prognosis,[18] and white blood cells (WBC) and monocytes (MONO) have also been reported to be related to cancers. All these markers can be measured from a single blood draw using routine methods available in clinical biochemistry/hematology laboratories. However, each of these markers is not specific to CRC when taken individually since many diseases and conditions can make them vary. Hence, global patterns of changes need to be determined to increase their diagnostic yield for CRC.

Machine learning, in which computers learn to generate their decision-making algorithms, was recently used in the diagnosis and prediction of CRC.[19–21] The power of this method lies in its ability to automatically make classification and identify primary features from millions of data and complex relationships. Lin et al.[23] developed a supervised random forest model to identify gene panels differentiating adenoma from CRC. Zhi et al.[21] used a support vector machine (SVM) classifier for the prediction of the metastasis of CRC. Li et al.[24] used an improved k-nearest neighbors classifier to diagnose CRC and colitis with Fourier transform infrared spectroscopy. Long et al.[25] examined the random forest, logistic regression, naïve Bayes, and k-nearest neighbors models and multi-platform transcriptomics to introduce novel signatures for the accurate diagnosis of CRC. Their results demonstrated that machine learning classifiers, including random forest, logistic regression, SVM, naïve Bayes, and k-nearest neighbors methods, could achieve a near-perfect accurate clinical diagnosis.[25]

Nevertheless, such models can always be improved by feeding them different variables. Therefore, this study aimed to determine models for CRC identification that involve minimally invasive, affordable, portable, and accurate screening variables, including laboratory detection data. Because they are being studied in medical fields and are showing good prospects,[19–24] the algorithms presented above were selected in the present study.

Materials and Methods

Dataset

This retrospective study used data from patients with CRC and healthy individuals who visited hospital between July 2017 and June 2018. All data were retrieved from the electronic medical record (EMR) system. The study was approved by the Ethics Committee of the Sixth Affiliated Hospital, Sun Yat-sen University (approval number: 2020ZSLYEL-081, approval date: May 8, 2020). Individual informed consent was waived due to the retrospective nature of this study.

Before feature extraction, the collected EMRs were deidentified after applying the inclusion and exclusion criteria. The inclusion criteria for the CRC patients were(1) the first diagnosis of CRC, without a history of any cancer treatment and (2) 30 to 70 years of age. The exclusion criteria were (1) a previous

history of cancer or CRC with a concurrent primary tumor at another site, (2) more than five comorbidities besides CRC, or (3) missing data.

For the healthy controls, the inclusion criteria were (1) 30 to 70 years of age and (2) underwent a general medical examination in our hospital. Those with incomplete data were excluded. The controls were matched 1:1 with the CRC patients for age and sex.

The EMR includes demographics, hospitalization information, biomedical test results, vital signs, diagnoses, and treatments for each patient. We extracted the features of interest from each of them to construct the dataset. According to the Guidelines of the Chinese Society of Clinical Oncology Standard, CRC was diagnosed based on colonoscopy and histopathologic biopsy.[26] The TNM stage of CRC was determined and was divided into early-stage (TNM stage of 0, I, and II) and late-stage (TNM stage of III and IV).

Data Collection

Twenty one input features were extracted from the EMR of each patient: three liver enzymes (ALT, AST, and GGT), seven lipid profiles (triglycerides [TG], total cholesterol [TC], HDL, LDL, apolipoprotein A1 [ApoA1], apolipoprotein B [ApoB], and Lp(a)), one inflammatory factor (high-sensitivity CRP [hs-CRP]), eight complete blood count parameters (RBC, HGB, WBC, NEU, LYM, MONO, ESO and PLT) and two tumor markers (CEA and AFP). The output feature was the confirmed clinical diagnosis. The vital signs were routinely collected on the first day of admission, and biomedical test results were routinely obtained on the second day of admission. The records of different patients did not vary by more than 2 days.

Feature Selection

To find the exact correlated variables in this study, we analyzed the relationships between every two variables by Spearman correlation analysis. For the variable pairs in which correlation coefficients were ≥0.50, the one that had the smallest weight coefficient in the principal component analysis was deleted from feature collection. All features were log-transformed to normalize their distribution and avoid within-subject differences in amplitude and variation among features.

Model Construction

Five machine learning models (eg, logistic regression, random forest, k-nearest neighbors, SVM, and naïve Bayes) were used to identify CRC. The machine learning algorithms for
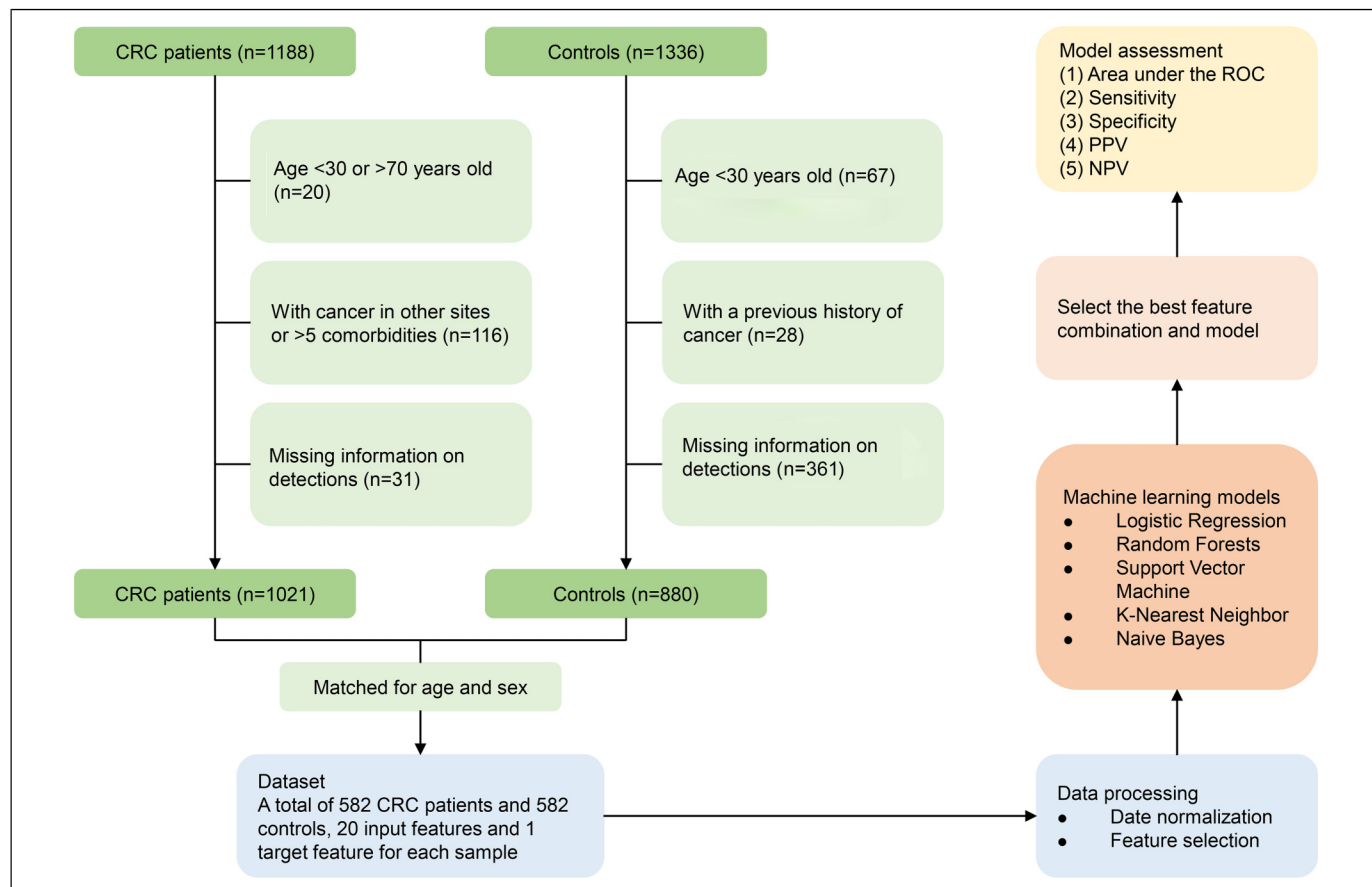

Figure 1. Flowchart of the colorectal cancer (CRC) identification model.

identification were run using a python library "Sklearn".[27] An L1 or L2 penalty was added as a hyperparameter to the logistic regression function and SVM function to reduce the risk of overfitting. An additional C parameter (10^{-4}-10^5 at 10 intervals) was required to control model complexity. Because the dataset was small and to avoid possible bias, the model's performance was assessed using stratified 10-fold cross-validation. Hence, 10 subsets were constructed by dividing the overall dataset randomly. Each of the 10 subsets was subsequently set as the testing set, with the remaining nine subsets as training sets. Then, the average of all 10 replicates was used to evaluate the performance of the models. The performances were evaluated by areas under the curve (AUCs), sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV).

Statistical Analysis

Continuous variables were tested for normal distribution using the Kolmogorov–Smirnov test. Those with a normal distribution were expressed as means ± standard deviation (SD) and analyzed using the Student's *t*-test. Those with a non-normal distribution were expressed as medians (Q1, Q3) and analyzed using the Mann–Whitney U test. Categorical variables were expressed as frequencies and analyzed using the chi-square test. *P*-values <.05 were considered statistically significant. Data were processed and analyzed using SPSS Modeler 18.1 (IBM). The machine learning models were trained and tested in Anaconda 3 (Anaconda Inc.).

Results

Subjects Screening

Initially, 1188 CRC patients were screened; 1168 patients met the inclusion criteria (20 patients were not aged 30 to 70 years). Then, 18 patients were excluded due to a previous history of cancer, 98 patients were excluded due to >5 comorbidities besides CRC, and 31 patients were excluded due to missing data. Finally, 1021 CRC patients were eligible for data analysis. A total of 1336 potential controls undergoing physical examination were screened; 1269 subjects met the inclusion criteria (67 subjects were not aged 30-70 years). Then, 28 subjects were excluded due to a previous history of cancer, and 361 were excluded due to missing data. Finally, 880 healthy subjects were included. After matching with age and sex, there were 582 pairs (Figure 1).

Characteristics of the Subjects

The baseline characteristics are presented in Table 1. CRC patients and healthy controls were matched for age and sex (52 ± 9 years old and 61% men). In CRC patients, 101 (17.4%) had early colon cancer, 164 (28.2%) had late colon cancer, 102 (17.5%) had early rectal cancer, and 215 (36.9%) had late rectal cancer.

Table 1. Characteristics of the Patients.

Baseline characteristic	CRCs (n = 582)	Controls (n = 582)
Sex (male)	355 (61%)	355 (61%)
Age (years)	52 ± 9	52 ± 9
CEA (ng/mL)	4.61 (2.38-13.59)*	1.70 (1.13-2.59)
α-fetoprotein (ng/mL)	2.52 (1.93-3.36)	3.49 (2.85-4.16)
Alanine transaminase (U/L)	13.76 (9.97-18.83)**	20.72 (15.57-29.69)
Aspartate transaminase (U/L)	17.02 (13.96-21.44)*	20.66 (17.64-24.72)
γ-glutamyltransferase (U/L)	20.93 (14.84-31.47)*	27.23 (17.50-42.98)
Triglycerides (mmol/L)	1.18 (0.92-1.58) **	1.45 (1.05-2.14)
Total cholesterol (mmol/L)	5.04 ± 1.12**	5.43 ± 1.05
HDL (mmol/L)	1.19 ± 0.30**	1.36 ± 0.31
LDL (mmol/L)	3.29 ± 0.83**	3.63 ± 0.82
ApoA1 (g/L)	1.16 ± 0.21**	1.35 ± 0.20
ApoB (g/L)	1.00 ± 0.24	0.90 ± 0.24
Lipoprotein (a) (g/L)	176.49 (101.62-342.76)**	103.15 (52.31-206.46)
hs-CRP (mg/L)	2.27 (0.82-1.13)**	0.88 (0.49-1.86)
Red blood cells (10^{12}/L)	4.50 ± 0.62**	4.86 ± 0.52
Hemoglobin (g/L)	120.64 ± 22.03**	141.56 ± 15.14
White blood cells (10^9/L)	6.76 ± 2.32**	6.02 ± 1.56
Neutrophils (10^9/L)	4.26 ± 2.03**	3.44 ± 1.14
Lymphocytes (10^9/L)	1.72 ± 0.63**	1.92 ± 0.55
Monocytes (10^9/L)	0.54 ± 0.23**	0.45 ± 0.15
Eosinophiles (10^9/L)	0.20 ± 0.17	0.18 ± 0.15
Platelets (10^9/L)	273.02 ± 97.58**	233.13 ± 57.25
Early colon cancer	101 (17.4%)	-
Late colon cancer	164 (28.2%)	-
Early rectal cancer	102 (17.5%)	-
Late rectal cancer	215 (36.9%)	-

Abbreviations: CRC: colorectal cancer; CEA: carcinoembryonic antigen; HDL: high-density lipoprotein; LDL: low-density lipoprotein; hs-CRP: high-sensitivity C-reactive protein.
*P < .05 versus the control group, **P < .001 versus the control group.

CEA levels were higher in CRC patients than in the controls (median: 4.61 vs 1.70 ng/mL, P = .006). All three liver enzymes of CRC patients were lower than the control group but within the normal ranges (ALT: P < .001; AST: P = .003; GGT: P = .048). HGB and RBC counts were significantly lower (both P <.001), while inflammation-associated markers, including WBC, NEU, MONO, and hs-CRP, were markedly increased in patients with CRC (all P < .001). TG, TC, HDL, LDL, and ApoA1 levels were lower in patients with CRC than in controls, while Lp(a) levels were higher (all P < .001). AFP, ApoB, and ESO were excluded from the dataset for no statistical significance between the two groups.

Correlation Analysis and Component Selection

The correlation coefficients among the 18 markers are presented in Figure 2. To resolve the multicollinearity, TC, AST, ApoA1, WBC, RBC, and MONO were excluded from the dataset since these markers had correlation coefficients >0.50. Finally, the components were selected for all five models: ALT, GGT, LDL, HDL, TG, hs-CRP, Lp(a), CEA, NEU, LYM, HGB, and PLT.
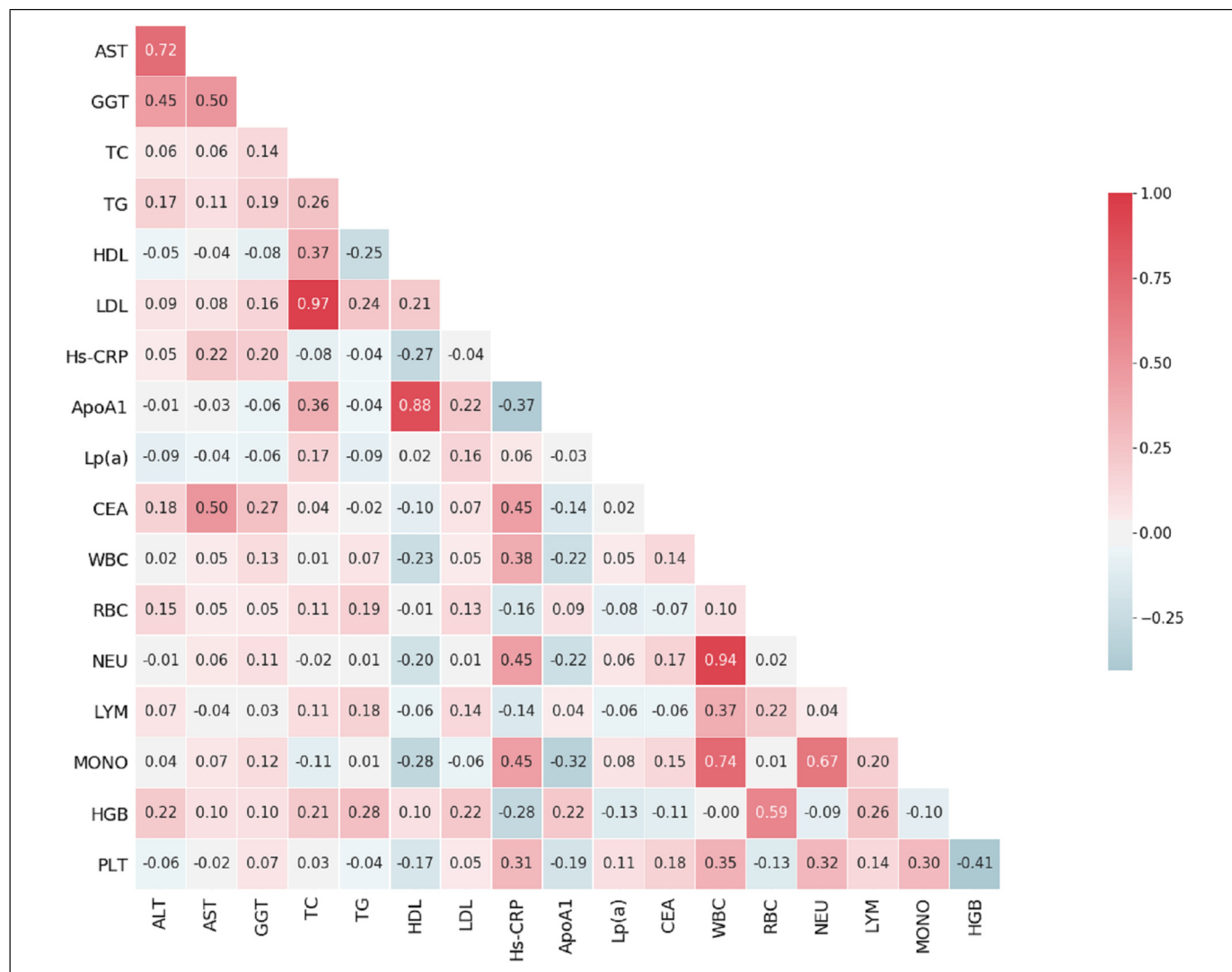
Figure 2. Matrix of the Spearman correlation coefficients. All pairs of variables included in the models were tested using the Spearman correlation. For the variable pairs in which correlation coefficients >0.5, the one with the less weight coefficient in the principal component analysis (PCA) was deleted from feature collection.

Model Selection

AUC was used to assess the performance of classification in five machine learning models. The logistic regression model and the SVM model had the highest diagnostic values among the five machine learning models to discriminate patients with CRC from healthy people, with AUCs of 0.865 (95% confidence interval [CI]: 0.857-0.877) and 0.865 (95% CI: 0.857-0.874), respectively. The logistic regression model had a mean sensitivity of 89.5% (95% CI: 88.0%-90.3%), slightly lower than that of the SVM model, whose mean sensitivity was 90.1% (95% CI: 88.0%-91.4%) (Table 2). The logistic regression model can indicate the importance of each variable and thus is more interpretable than the SVM model in clinical

Table 2. Performance Comparison of Dfferent Machine Learning Models in Colorectal Cancer (CRC) (vs health).

Patterns	AUC	Sensitivity	Specificity	PPV	NPV
Logistic Regression	0.865 (0.857- 0.877)	0.895 (0.880-0.903)	0.835 (0.817-0.851)	0.844 (0.835-0.859)	0.889 (0.875-0.898)
Random Forests	0.848 (0.840-0.857)	0.873 (0.857-0.891)	0.823 (0.80-0.840)	0.832 (0.817 to 0.848)	0.867(0.852-0.885)
Support Vector Machine	0.865 (0.857-0.874)	0.901 (0.880-0.914)	0.830 (0.806-0.851)	0.842 (0.827-0.859)	0.894 (0.879-0.903)
K-Nearest Neighbor	0.816 (0.797-0.831)	0.879 (0.851-0.897)	0.754 (0.737-0.771)	0.781 (0.771-0.796)	0.863 (0.830-0.883)
Naive Bayes	0.839 (0.831-0.849)	0.928 (0.914-0.943)	0.749 (0.731-0.766)	0.788 (0.777-0.796)	0.913 (0.903-0.923)

Abbreviations: CEA: carcinoembryonic antigen; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value.
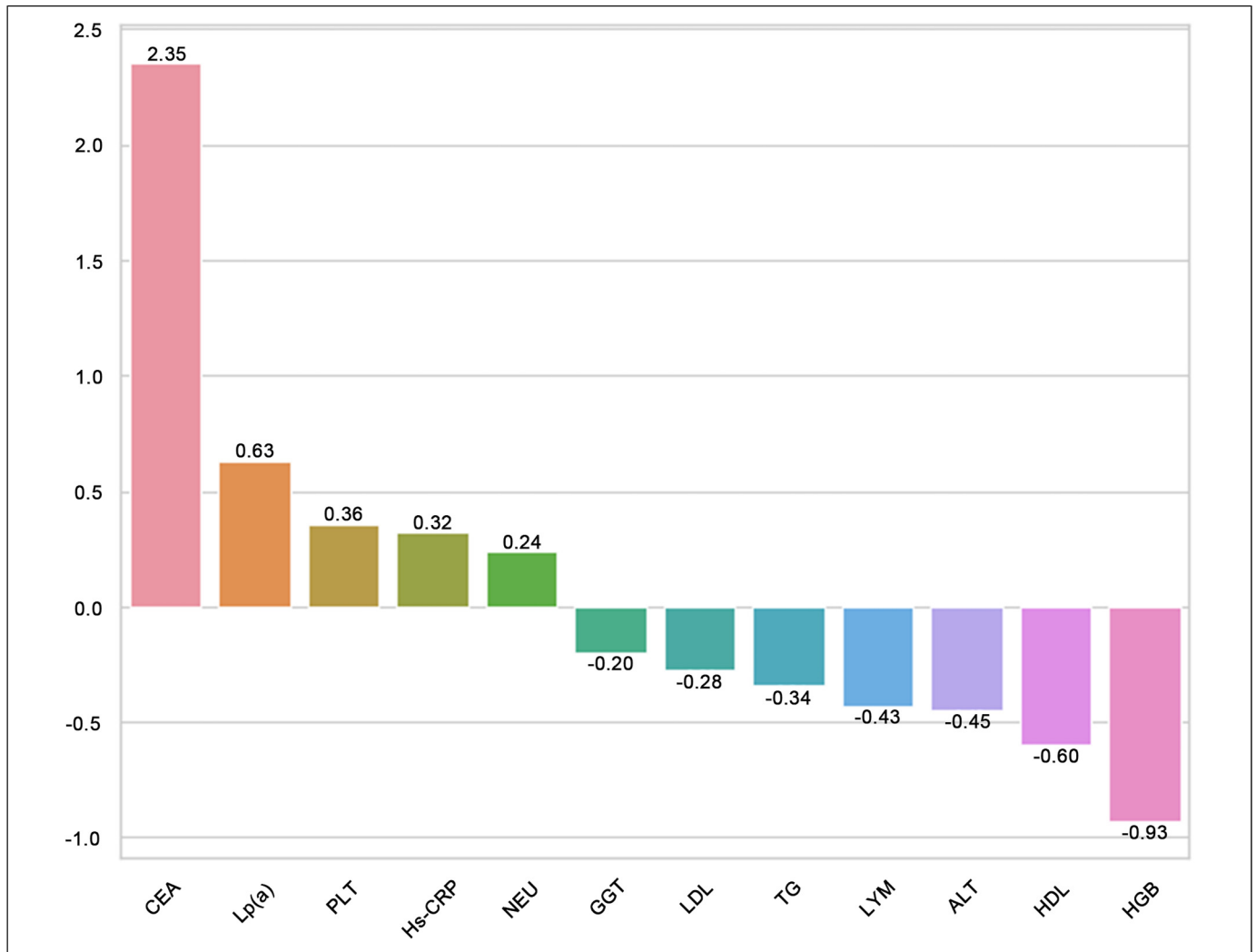
Figure 3. The weight coefficients of the logistic regression model (the model with the highest accuracy) for colorectal cancer (CRC) diagnosis. The first four weighted features in the logistic regression model were carcinoembryonic antigen (CEA), hemoglobin (HGB), lipoprotein (a) (Lp(a)), and high-density lipoprotein (HDL).

practice; it is chosen as the newly developed model in this study.

In this logistic regression model, CEA, HGB Lp(a), and HDL were of the utmost importance in CRC discrimination, with a weight of 2.35, −0.93, 0.63, and −0.60, respectively (Figure 3). A diagnostic model for CRC was established based on the four indicators, considering that it achieved a similar AUC with that of the model using a combination of all the 12 parameters (0.849 and 0.865, respectively, Figure 4).

Classification Performance

The proposed model discriminated CRC patients from healthy individuals with an AUC of 0.849 (0.840 to 0.860), a sensitivity of 88.3% (87.4% to 90.3%), and a specificity of 81.5% (79.4%-83.4%). This model had better performance for diagnosing colon cancer than for rectal cancer, and better

performance for diagnosing late-stage than for early-stage CRC. This model had the best AUC (0.905 [0.889-0.929]) in discriminating patients with late colon cancer from healthy individuals among CRC patients with different disease sites and stages (Table 3).

Discussion

This study aimed to develop and assess the performance of a machine learning system to identify CRC patients from healthy individuals using the most common markers in conventional laboratory blood tests. Our main finding was that a logistic regression algorithm, based on markers of CEA, HGB Lp(a), and HDL, could discriminate CRC patients, especially late colon patients, from healthy individuals. In our test sets, the sensitivity for detecting CRC was 88.3%, and the specificity was 81.5%.
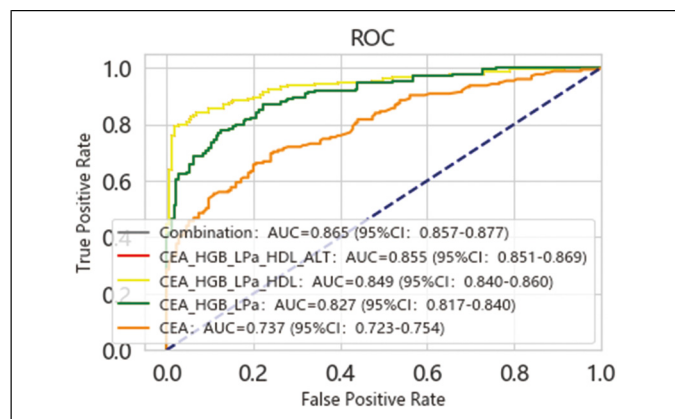
Figure 4. Receiver operating characteristic (ROC) curve for colorectal cancer (CRC) diagnosis using logistic regression models: CEA alone, CEA + hemoglobin (HGB) + Lp(a), CEA + HGB + Lp(a) + HDL, and CEA + HGB + Lp(a) + HDL + ALT.
Abbreviations: ALT: alanine transaminase; CEA: carcinoembryonic antigen; HDL: high-density lipoprotein; Lp(a): lipoprotein (a).

The comparison of the models studied here with the models from other studies is challenging because the majority of the studies on CRC focused on CRC identification using cutting-edged markers, which is different from the purpose of the present study, which was to use markers commonly tested in routine medical practice. Therefore, other studies included different variables and tested different models, but such a comparison still has some value. Table 4[19,20,22,25,28,29] presents the results from previous studies and those of the present one. AUC reflects the overall diagnostic efficiency of the models. Sensitivity indicates the probability that CRC cases are correctly diagnosed, specificity that the non-CRC individuals are correctly classified. PPV is the proportion of actual CRC patients in individuals diagnosed with CRC by the model, and NPV is the proportion of actual non-CRC individuals in those classified as non-CRC ones by the model. The molecular biological techniques obviously had higher accuracy in terms of AUC. Long et al.[25] found that the random forest model based on high-throughput omics was the best method for prediction tasks and achieved the highest AUC of 0.998. Nakajima et al.[20] used the ADTree model, which is a kind of decision tree based on a boosting algorithm, to develop the prediction model and obtained good predictive value than others, but only 59 patients were included, which is small for a validation study. The next two studies both focused on the early prediction of CRC. In the study by Wan et al.,[22] the logistic regression combined vector machine was used in whole-genome sequencing of plasma cell-free DNA; the AUC and sensitivity were similar to the proposed model. Hornbrook et al.[19] used sex, age, and complete blood count data to identify early-stage CRC; the AUC was 0.80, which is not optimal. Kinar et al.[28] used complete blood count data and achieved an AUC of 0.82, close to that of Hornbrook et al..[19] Zhao et al.[29] used age, body mass index (BMI), and gut bacteria to achieve an AUC of 0.942.

One of the confusing factors in artificial intelligence is that even though the variables are selected based on a clinical rationale, the exact clinical significance of the results cannot be explained clearly most of the time. The explanation can only

Table 3. Performance Comparison of Different Colorectal Cancer Patterns (vs health) Based on CEA, Hemoglobin, HDL, and Lp(a) (logistic regression model).

Patterns	AUC	Sensitivity	Specificity	PPV	NPV
All CRC	0.849 (0.840-0.860)	0.883 (0.874-0.903)	0.815 (0.794 to 0.834)	0.828 (0.815-0.840)	0.875 (0.866-0.888)
Early colon cancer	0.801 (0.785-0.820)	0.859 (0.833-0.900)	0.743 (0.710- 0.774)	0.771 (0.750-0.784)	0.844 (0.821-0.880)
Late colon cancer	0.905 (0.889-0.929)	0.921 (0.898-0.959)	0.888 (0.878-0.898)	0.892 (0.880-0.906)	0.920 (0.896-0.956)
Early rectal cancer	0.745 (0.742-0.758)	0.792 (0.774-0.839)	0.698 (0.645-0.742)	0.728 (0.711-0.758)	0.773 (0.739-0.815)
Late rectal cancer	0.862 (0.852-0.883)	0.886 (0.862-0.908)	0.838 (0.800-0.877)	0.849 (0.819-0.875)	0.881 (0.857-0.898)

Abbreviations: CRC: colorectal cancer; CEA: carcinoembryonic antigen; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value.

Table 4. Comparison of Performance Between the Proposed Model and From Other Studies.

Author	Detections	Algorithms	AUC	Sensitivity	Specificity
Long[25]	Multi-platform transcriptomics	RF	0.998 (0.995-0.999)	99.8%	99.9%
Nakajima[20]	Urinary polyamine biomarker panel	ADTree	0.961 (0.937-0.984)	N/A	N/A
Wan[22]	Whole-genome sequencing	LR + SVM	0.92 (0.91 to 0.93)	85%	85%
Hornbrook[19]	Complete blood count	ColonFlag®	0.80 (0.79-0.81)	N/A	N/A
Kinar[28]	Complete blood counts	GBM + RF	0.82	50%	87%
Zhao[29]	Age, BMI, gut bacteria	LR + SVM	0.942	93.3%	80.7%
Proposed	CEA, hemoglobin, HDL, and Lp(a)	LR	0.849 (0.840-0.860)	88.3%	81.5%

Abbreviations: ADTree: alternating decision tree; AUC: area under the curve; BP: backpropagation; CEA: carcinoembryonic antigen; GBM: gradient boosting model; LR, logistic regression; RF, random forests; SVM, support vector machine.

be interpreted in the light of the complex interrelationships among the variables determined by the machine learning methods. In this model, the relative importance of each feature could be explained by their regression coefficient. The first four weighted features in the model were CEA, HGB, Lp(a), and HDL. Those variables are already known to be associated with CRC,[6–12] suggesting the clinical robustness of the model. In addition, these markers are easy to obtain in a clinical setting through a simple blood draw, without invasive examinations and expensive costs.

The model appears to have better accuracy for colon cancer than for rectal cancer and for late-stage than early-stage CRC. More colon cancers are diagnosed each year than rectal cancer,[30] which might bias the results, and no firm conclusion can be drawn for the moment. On the other hand, regarding the stage at diagnosis, late-stage cancers are associated with more pronounced metabolic disturbances, particularly in the presence of metastasis, than early-stage cancer, making late-stage cancers easier to detect, as seen in screening studies.[31] Still, the AUCs, sensitivity, and specificity remain good for early CRC, suggesting the value of using the simple model based on clinical blood markers.

The biochemical and hematological markers included in the models are all tests that are available in clinical laboratories. Still, some of the features in the models are available but not routinely examined, eg, Lp(a). Thus, the results may suggest that the clinicians should include these indicators as routine blood tests when considering CRC. In addition, there are many clinical features that may suggest CRC, and the aim of this study was to establish a predictive model that could identify CRC with as few variables as possible. This would allow the patients to undergo only a small No. of tests initially and undergo further tests when they are at relatively high risk of CRC. The results suggest that the physicians should pay attention to the biomarkers included in the model when considering a suspicion of CRC. With the high sensitivity and AUC, the use of the model could confirm the physician's suspicion and prompt additional examinations. Still, this model has to be validated in a larger population and has to be examined in a screening context.

This study has limitations. First, because of the selection criteria, the analyses were performed in selected patients, and the results might not apply to all patients with CRC or early CRC, particularly if the patients have another cancer or a benign tumor in addition to CRC. Second, the patients were from July 2017 to June 2018 and from a single center, which may limit the generalization ability of the conclusions. Thus, the applicability of the findings to other hospitals remains to be determined. Third, the inclusion of two controls for each patient has been suggested to be the best approach,[32] but it was not possible in the present study since there were 880 healthy controls and 1021 CRC patients eligible for this study before matching. Future prospective studies with larger sample size are needed to verify the results.

In conclusion, the present study strongly suggests that logistic regression is a powerful tool to identify patients with CRC.

This study also provides a feature set for the task of identifying CRC using machine learning approaches. That is promising to provide decision support for clinicians in the presence of a complicated case. The logistic regression model based on conventional laboratory test data could be a powerful, noninvasive, and cost-effective method to identify CRC, especially late-stage colon cancer.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Approval

This study was approved by the Ethics Committee of the Sixth Affiliated Hospital, Sun Yat-sen University (approval number: 2020ZSLYEL-081, approval date: May 8, 2020). Individual informed consent was waived due to the retrospective nature of this study.

ORCID iD

Huanliang Liu (iD) https://orcid.org/0000-0002-1006-6666

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(3):209-249.
2. Crockett SD, Nagtegaal ID. Terminology, molecular features, epidemiology, and management of serrated colorectal neoplasia. *Gastroenterology.* 2019;157(4):949-966.
3. He N, Song L, Kang Q, et al. The pathological features of colorectal cancer determine the detection performance on blood ctDNA. *Technol Cancer Res Treat.* 2018;17:1–9.
4. Bénard F, Barkun AN, Martel M, et al. Systematic review of colorectal cancer screening guidelines for average-risk adults: summarizing the current global recommendations. *World J Gastroenterol.* 2018;24(1):124-138.
5. Becker D, Grapendorf J, Greving H, et al. Perceived threat and internet Use predict intentions to Get bowel cancer screening (colonoscopy): longitudinal questionnaire study. *J Med Internet Res.* 2018;20(2):e46.

6. Hundt S, Haug U, Brenner H. Blood markers for early detection of colorectal cancer: a systematic review. *Cancer Epidemiol Biomarkers Prev*. 2007;16(10):1935-1953.

7. Thomas DS, Fourkala EO, Apostolidou S, et al. Evaluation of serum CEA, CYFRA21–1 and CA125 for the early detection of colorectal cancer using longitudinal preclinical samples. *Br J Cancer*. 2015;113(2):268-274.

8. Väyrynen JP, Tuomisto A, Väyrynen SA, et al. Preoperative anemia in colorectal cancer: relationships with tumor characteristics, systemic inflammation, and survival. *Sci Rep*. 2018;8(1): 1126.

9. van Duijnhoven FJB, Bueno-De-Mesquita HB, Calligaro M, et al. Blood lipid and lipoprotein concentrations and colorectal cancer risk in the european prospective investigation into cancer and nutrition. *Gut*. 2011;60(8):1094-1102.

10. Zhang X, Zhao X-W, Liu D-B, et al. Lipid levels in serum and cancerous tissues of colorectal cancer patients. *World J Gastroenterol*. 2014;20(26):8646-8652.

11. Pakiet A, Kobiela J, Stepnowski P, Sledzinski T, Mika A. Changes in lipids composition and metabolism in colorectal cancer: a review. *Lipids Health Dis*. 2019;18(1):29.

12. Katzke VA, Sookthai D, Johnson T, Kühn T, Kaaks R. Blood lipids and lipoproteins in relation to incidence and mortality risks for CVD and cancer in the prospective EPIC-Heidelberg cohort. *BMC Med*. 2017;15(1):218.

13. Wu X-Z, Ma F, Wang X-L. Serological diagnostic factors for liver metastasis in patients with colorectal cancer. *World J Gastroenterol*. 2010;16(32):4084-4088.

14. Li JY, Li Y, Jiang Z, Wang RT, Wang XS. Elevated mean platelet volume is associated with presence of colon cancer. *Asian Pac J Cancer Prev*. 2014;15(23):10501-10504.

15. Li L, Huang X-Y, Li N, Cui M-M, Wang R-T. Platelet indices in colorectal cancer patients with synchronous liver metastases. *Gastroenterol Res Pract*. 2019;2019:6397513.

16. Greten FR, Grivennikov SI. Inflammation and cancer: triggers, mechanisms, and consequences. *Immunity*. 2019;51(1):27-41.

17. Goyal A, Terry MB, Jin Z, Siegel AB. C-reactive protein and colorectal cancer mortality in U.S. Adults. *Cancer Epidemiol Biomarkers Prev*. 2014;23(8):1609-1618.

18. Mazaki J, Katsumata K, Kasahara K, et al. Neutrophil-to-lymphocyte ratio is a prognostic factor for colon cancer: a propensity score analysis. *BMC Cancer*. 2020;20(1):922.

19. Hornbrook MC, Goshen R, Choman E, et al. Early colorectal cancer detected by machine learning model using gender, Age, and complete blood count data. *Dig Dis Sci*. 2017;62(10):2719-2727.

20. Nakajima T, Katsumata K, Kuwabara H, et al. Urinary polyamine biomarker panels with machine-learning differentiated colorectal cancers, benign disease, and healthy controls. *Int J Mol Sci*. 2018;19(3):756.

21. Zhi J, Sun J, Wang Z, Ding W. Support vector machine classifier for prediction of the metastasis of colorectal cancer. *Int J Mol Med*. 2018;41(3):1419-1426.

22. Wan N, Weinberg D, Liu T-Y, et al. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer*. 2019;19(1):832.

23. Lin S-H, Raju GS, Huff C, et al. The somatic mutation landscape of premalignant colorectal adenoma. *Gut*. 2018;67(7):1299-1305.

24. Li Q, Hao C, Kang X, et al. Colorectal cancer and colitis diagnosis using Fourier transform infrared spectroscopy and an improved K-nearest-neighbour classifier. *Sensors (Basel)*. 2017;17(12):2739.

25. Long NP, Park S, Anh NH, et al. High-Throughput omics and statistical learning integration for the discovery and validation of novel diagnostic signatures in colorectal cancer. *Int J Mol Sci*. 2019;20(2):296.

26. Diagnosis and Treatment Guidelines For Colorectal Cancer Working Group C. Chinese Society of clinical oncology (CSCO) diagnosis and treatment guidelines for colorectal cancer 2018 (English version). *Chinese Journal of Cancer Research = Chung-kuo yen Cheng yen Chiu*. 2019;31(1):117-134.

27. Fabian P, Gaël V, Alexandre G, et al. Scikit-Learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.

28. Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *J Am Med Inform Assoc*. 2016;23(5):879-890.

29. Zhao D, Liu H, Zheng Y, et al. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Med Biol Eng Comput*. 2019;57(4):901–912.

30. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin*. 2020;70(3):145–164.

31. Andrew AS, Parker S, Anderson JC, et al. Risk Factors for Diagnosis of Colorectal Cancer at a Late Stage: a Population-Based Study. *J Gen Intern Med*. 2018;33(12):2100–2105.

32. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics*. 1975;31(3):643–649.