Research Article

Explainable Machine Learning Models for Colorectal Cancer Prediction Using Clinical Laboratory Data

Cancer Control
Volume 32: 1–14

DOI: 10.1177/10732748251336417
journals.sagepub.com/home/ccx
Sage

Rui Li, MS[1,†], Xiaoyan Hao, MS[1,†], Yanjun Diao, MD[1,†], Liu Yang, MS[1], and Jiayun Liu, MD[1]

Abstract

Introduction: Early diagnosis of colorectal cancer (CRC) poses a significant clinical challenge. This study aims to develop machine learning (ML) models for CRC risk prediction using clinical laboratory data.

Methods: This retrospective, single-center study analyzed laboratory examination data from healthy controls (HC), polyp patients (Polyp), and CRC patients between 2013 and 2023. Five ML algorithms, including adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), decision tree (DT), logistic regression (LR), and random forest (RF), were employed to classify subjects into HC vs Polyp vs CRC, HC vs CRC, and Polyp vs CRC, respectively.

Results: This study included 31 539 subjects: 11 793 HCs, 10 125 polyp patients, and 9621 CRC patients. The XGBoost model achieved the highest AUCs of 0.966 for differentiating HC from CRC and 0.881 for Polyp from CRC, outperforming carcino-embryonic antigen (CEA) and fecal occult blood testing (FOBT) tests. This model could also identify CEA-negative or FOBT-negative CRC patients. Incorporating stool miR-92a detection into the model further improved diagnostic performance. Shapley additive explanations (SHAP) plots indicated that FOBT, CEA, lymphocyte percentage (LYMPH%), and hematocrit (HCT) were the most significant features contributing to CRC diagnosis. Additionally, a computational tool for predicting CRC risk based on the optimal model was developed, designed for researchers with programming experience.

Conclusion: Five ML models for CRC diagnosis, based on ten routine laboratory test items, were developed, achieving higher diagnostic accuracies than traditional CRC biomarkers. The diagnostic capabilities of these ML models can be further enhanced by including stool miR-92a levels.

Keywords

colorectal cancer, machine learning, clinical laboratory data, risk prediction, miR-92a

Received: 25 January 2025; revised: 13 March 2025; accepted: 1 April 2025.

[1]Department of Clinical Laboratory Medicine, Xijing Hospital, Air Force Medical University, Xi'an, China

[†]These authors contributed equally to this work and share the first authorship

Corresponding Authors:
Liu Yang, Department of Clinical Laboratory Medicine, Xijing Hospital, Air Force Medical University, No. 127, Changle West Road, Xincheng District, Xi'an 710032, China.
Emails: zhangzh@fmmu.edu.cn; yangliuxjyy@126.com

Jiayun Liu, Department of Clinical Laboratory Medicine, Xijing Hospital, Air Force Medical University, No. 127, Changle West Road, Xincheng District, Xi'an 710032, China.
Email: jiayun@fmmu.edu.cn

Introduction

Colorectal cancer (CRC) is a leading global health concern, ranking as the third most common cancer and the second leading cause of cancer-related deaths worldwide.[1] CRC survival rate is significantly associated with the clinical stage at diagnosis, with a 5-year survival rate of over 90% for stage I patients, but falling below 25% for those diagnosed at stage IV.[2] Early diagnosis of CRC may lead to prompt treatment, which can substantially improve outcomes.[3,4] Several methods have been developed to identify CRC patients at early stages, including colonoscopy,[5] computed tomography (CT),[6] fecal DNA testing,[7] fecal occult blood testing (FOBT),[8] carcinoembryonic antigen (CEA),[9] and other tumor biomarkers.[10] However, these methods are often limited because they are either highly invasive, or show insufficient specificity and sensitivity.[11-14] Recent advancements in medical omics research have spurred the rapid development of early screening strategies for CRC, based on metabolomics and genomics. Innovative technologies, such as circulating tumor DNA (ctDNA)[15] testing, fecal microRNA (miRNA) analysis, fecal multi-target DNA (FIT-DNA) testing,[11] and metabolomics methods—including ultra-performance liquid chromatography-mass spectrometry (UPLC-MS),[16] nuclear magnetic resonance spectroscopy (NMR),[17] and gas chromatography time-of-flight mass spectrometry (GC-TOFMS)[18]—offer new possibilities for early CRC detection. These approaches analyze genetic variations, miRNA expression, and metabolic changes to improve screening accuracy. However, they also face significant limitations, including high costs, expensive equipment, and the need for skilled professionals, making widespread implementation challenging. Despite the effectiveness of CRC screening, it remains underutilized owing to potential risks and high costs, necessitating the need for novel screening methods with improved accessibility and cost-effectiveness.

CRC onset and progression are complex physiological and pathological processes encompassing gene mutations,[19] hypoxic microenvironments,[20] and immune escape.[21] During these processes, cancer cells, somatic cells, and immune cells interact with each other in a highly dynamic and heterogeneous manner. This complexity makes it extremely difficult to detect and monitor CRC by measuring only one or a few tumor markers, especially in the early stages where the progression is often slow and subtle. One potential solution to this challenge is the rational combination of a large number of biomarkers to capture a more comprehensive picture of CRC development and progression, thereby improving diagnostic accuracy.[22] Routine laboratory parameters, which are readily available in clinical settings, have been extensively used to diagnose, exclude, or monitor various diseases due to their sufficient validity and stability across large-scale evaluations. However, with over 1000 laboratory testing items, it is impractical to manually identify which parameters are associated with CRC development, let alone leverage them for accurate CRC diagnosis.

Machine learning (ML) has emerged as a powerful tool in data analysis and pattern recognition in healthcare.[23] By training ML models on large, labelled clinical datasets, these models can identify hidden but clinically relevant insights and predict patient outcomes, thereby supporting clinical decision-making and enhancing diagnostic accuracy.[24,25] Cancers, a complex and diverse group of prevalent diseases, present challenging diagnostic problems and generate extensive data across various modalities. This makes clinical oncology a fertile ground for the application of ML. ML techniques are employed to analyze medical imaging and molecular data derived from both liquid and solid tumor biopsies.[26] Specifically, ML algorithms, including deep learning models, are adept at scrutinizing these images to detect cancer-related patterns. ML models also excel at analyzing biomarkers such as cell-free DNA (cfDNA), methylation status, and gene expression profiles. The integration of multiple molecular features with ML enhances the sensitivity and specificity of cancer detection. In the United States, most ML algorithms are regulated as medical devices by the Food and Drug Administration (FDA). Over the past decade, the FDA has approved more than 300 AI/ML-enabled medical devices, with over 40% receiving approval since 2020.[27] As for CRC, the development of computer vision (CV) and convolutional neural network (CNN) methods have enabled the automatic detection of CRC from clinical images, such as colonoscopy images and histological whole slide images, showing great potential for future CRC diagnosis.[28,29] In this study, we aim to apply various ML algorithms to mine large-scale clinical laboratory data and identify key CRC-associated laboratory parameters, enabling accurate diagnosis at early stages. In addition, we further explored whether including fecal miR-92a, a microRNA reported to be elevated in CRC patients[30] and utilized for clinical CRC screening in certain countries, could enhance the diagnostic accuracy of the ML models.

Methods

Study Design and Population

This retrospective, single-center study utilized data from patients admitted to the First Affiliated Hospital of the Air Force Medical University between January 2013 and December 2023. All data were anonymized prior to analysis, and each research subject was assigned a unique identification number. We collected patients' baseline information, diagnoses, laboratory test results, colonoscopy findings, and pathological reports. All test results were gathered within 15 days prior to the colonoscopy examination or definitive diagnosis. This study was approved by the Medical Ethics Committee of the First Affiliated Hospital of the Air Force Medical University of the People's Liberation Army with the approval number KY20242119. Informed consent was waived

due to the study's retrospective design. The reporting of this study conforms to STROBE guidelines.[31] All patient details used in this study have been thoroughly de-identified to safeguard the privacy and confidentiality of the individuals involved.

For CRC patients, the inclusion criteria were: (1) confirmed CRC diagnosis based on pathology and (2) age between 30 and 70 years. The exclusion criteria were: (1) a history of other tumors, (2) prior anti-cancer treatments such as surgery, radiotherapy, chemotherapy, or immunotherapy, (3) a diagnosis of inflammatory bowel disease including ulcerative colitis or Crohn's disease, and (4) severe comorbidities (such as myocardial infarction, heart failure, cerebral hemorrhage, cerebral infarction, systemic lupus erythematosus and rheumatoid arthritis, diabetes mellitus, acute and chronic nephritis and renal failure, cirrhosis, abnormal coagulation and both acute and chronic leukemia) unrelated to CRC. For patients with colorectal polyps (all type of polyps), the inclusion criterion was a first-time diagnosis of colorectal polyps, with those having serious systemic diseases excluded. Healthy controls (HC) were included based on the following criteria: (1) age between 30 and 70 years, and (2) underwent a general medical examination with no major disease diagnosed. Cohort 1, which includes those from before 2023 focusing on feature extraction and model building, and Cohort 2, which consists of data collected since 2023 for prospective validation. More detailed information is displayed in Supplemental Table S1.

Data Cleaning and Standardization

After obtaining the original data, preprocessing was conducted. First, indicators that were missing in more than 30% of patients were removed from the dataset. For patients with multiple tests, only the results from the first test were recorded. The dataset was then standardized using the Logical Observation Identifiers Names and Codes (LOINC) system. Next, quantitative data were binarized, while qualitative data were encoded using dummy coding. Normalization is the process of scaling variables with values outside the 0-1 range to fit within that range. In this study, the MinMaxScaler from the sklearn. preprocessing package in Python was used to normalize the data:

$$v_{norm} = (v - V_{min})/(V_{max} - V_{min}),$$

where v represents the raw data, Vmin represents the minimum value of this index within the dataset, and Vmax represents the maximum value of this index within the dataset. Missing data were supplemented by the K-nearest neighbor (KNN) algorithm. Specifically, the distance matrix of the data points with missing values and other data points without missing values was calculated, the K data points with the closest Euclidean distance were selected, and the field mean of the corresponding data points of the selected k nearest neighbors were used to fill in the missing data values.

ML Algorithm and Model Building

After data preprocessing, meaningful features were selected to input into the ML algorithm for training. Features in cohort 1 were extracted using recursive feature elimination (RFE), Spearman correlation coefficients, and mutual information (MI). The intersection of the top 20 predictors extracted by each method was used as the final algorithm features to develop ML models. We established five common ML classifiers, including adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), decision tree (DT), logistic regression (LR), and random forest (RF), using data in cohort 1 using the sci-kit-learn (v.1.0.1) package. In Cohort 1, 90% of the data, consisting of 27 417 cases, was used for training, while 10%, or 3047 cases, was set aside for testing. For each algorithm, one model was trained to classify CRCs, HCs, and Polyps, and two additional models were trained to classify CRC vs HCs, and CRC vs Polyps, respectively. All classifiers were optimized using 10-fold cross-validation.

The performance of these models was then validated using data in Cohort 2 using the same features as those selected in the training process. When incorporating stool miR-92a for model establishment, the data were re-divided in a 4:1 ratio for training and validation. The trained models were also compared with the traditional CRC laboratory detection indicators CEA and FOBT. Furthermore, a subgroup analysis was conducted to test the model's ability to identify patients with FOBT-negative CRC and non-elevated CEA expression. Sensitivity, specificity, accuracy, positive predictive value, negative predictive value, and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve were used to evaluate the performance of the models. The data collection, pre-processing, and model construction process is illustrated in Supplemental Fig. S1. Shapley additive explanations (SHAP) was used to interpret variable importance.

Stool miR-92a Detection

Since 2023, all patients in Cohort 2 have received fecal miR-92a testing in addition to clinical laboratory tests. Those who underwent CRC surgery had a follow-up fecal miR-92a test one-month post-surgery. Specifically, stool (0.3-0.5 g) total RNA was first extracted using the REColon™ Nucleic Acid Extraction kit (GeneBioHealth, China) according to the manufacturer's instructions. After RNA extraction, miR-92a was quantified using the REColon TM miR-92a Quantitative PCR Kit (GeneBioHealth, Shenzhen, China) following the manufacturer's instructions. The detection methods for other indicators are detailed in the Supplemental Methods.

Statistical Analysis

Statistical analysis was conducted using Python version 3.8 (https://www.python.org/) and R software version 3.1.1 (https://www.r-project.org/). The specific libraries and their

versions are as follows: the Scikit - learn library version 1.3.2 was used for modeling operations; the numpy library version 1.22.4 was mainly used for numerical operations; the matplotlib library version 3.4.1 was used for data visualization; the scipy library version 1.6.3 assisted in scientific computing tasks; and the xlrd library version 1.2.0 was responsible for reading Excel files. Normally distributed continuous variables are presented as mean and standard deviation. Comparison between two groups was performed using Student's *t*-test. $P < .05$ were considered statistically significant.

Results

Baseline Characteristics and Feature Extraction

A total of 31 539 objects were included in the study, including 11 793 healthy controls, 10 125 polyp patients, and 9621 CRC patients (Figure 1). Cohort 1 and Cohort 2 included 30 464 and 1075 patients, respectively.

After data preprocessing and feature selection, 66 items were extracted from 1415 routine laboratory tests as candidate features for modeling, as most items had over 30% missing data. The top 20 indicators and their importance in the RFE, Spearman, and MI methods are listed in SupplementalTable S2. The intersection of these included 10 features: lymphocyte percentage (LYMPH%), hemoglobin (HGB), hematocrit (HCT), red blood cell distribution width (RDW%), platelet distribution width (PDW), total protein (TP), glucose (GLU), CEA, FOBT, and uric acid (UA) (Supplemental Fig. S2A). Besides the well-known items FOBT and CEA, the LYMPH %, a common indicator of routine blood tests, was the most important CRC prediction factor, followed by HCT, hemoglobin HGB, and RDW%. We further analyzed the correlation between these features using Spearman correlation coefficient and discovered that HCT and HGB had a strong positive correlation (correlation coefficient = 0.97). The correlation coefficients between other indicators were <0.6 (Supplemental Fig. S2B).

CRC Prediction Model Establishment

All five ML algorithms showed remarkable performance in distinguishing HCs, Polyps, and CRCs (Supplemental Table S3). The RF model showed the highest accuracy among all models in the training and test datasets, reaching 99.93% and

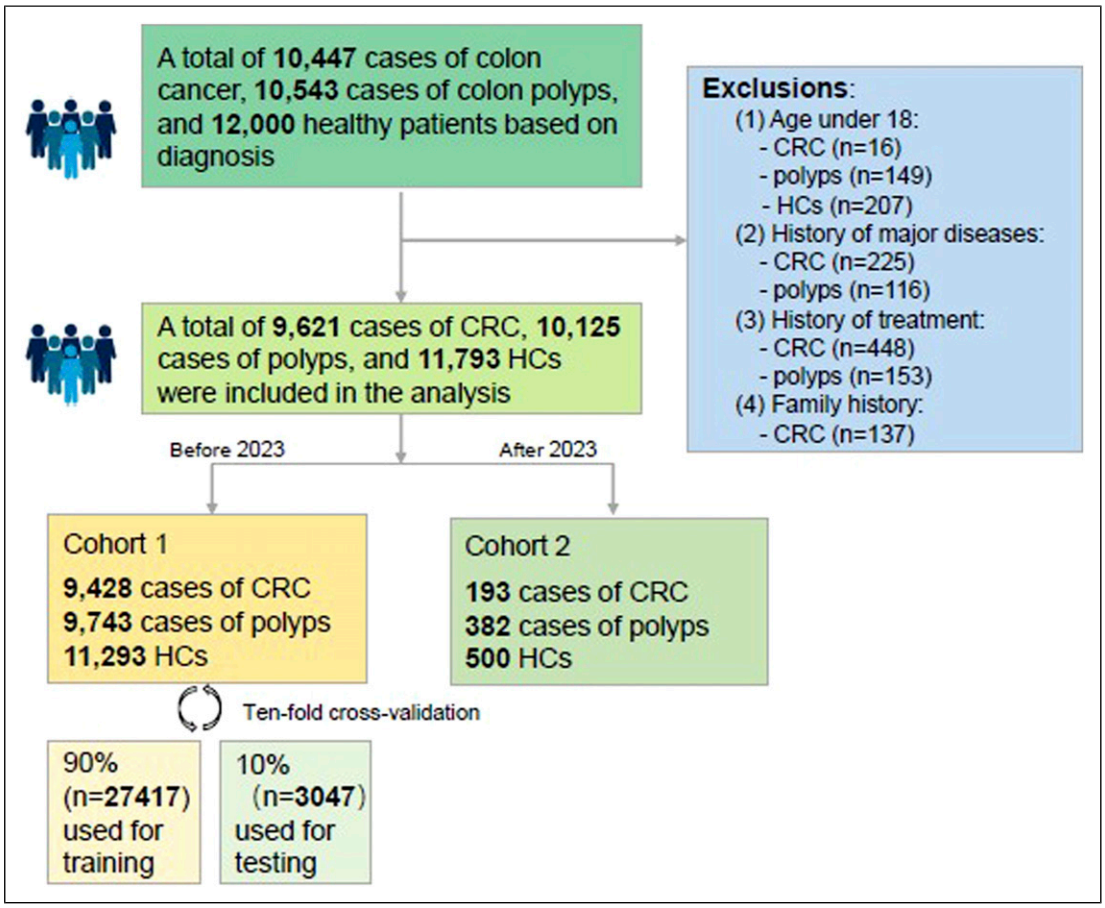

Figure 1. Study Design and Enrollment of Participants

Table 1. Performance of Five ML Models for Binary Classification.

	HC VS CRC						Polyp VS CRC					
Model	SEN	SPE	ACC	PPV	NPV	AUC	SEN	SPE	ACC	PPV	NPV	AUC
Training cohort												
AdaBoost	92.89%	97.07%	95.17%	96.36%	94.24%	0.983(0.981-0.985)	83.23%	88.22%	85.76%	87.24%	84.46%	0.924(0.920-0.928)
XGBoost	93.13%	98.17%	95.87%	97.70%	94.48%	0.985(0.983-0.987)	85.41%	89.76%	87.62%	88.98%	86.41%	0.940(0.936-0.944)
DT	89.93%	97.51%	94.06%	96.81%	92.07%	0.962(0.959-0.965)	80.95%	88.54%	84.81%	87.24%	82.78%	0.904(0.900-0.908)
LR	90.03%	97.77%	94.25%	97.12%	92.16%	0.978(0.976-0.980)	79.45%	88.54%	84.07%	87.03%	81.66%	0.903(0.899-0.907)
RF	100.00%	100.00%	100.00%	100.00%	100.00%	0.995(0.994-0.996)	99.92%	99.85%	99.89%	99.85%	99.93%	0.995(0.994-0.996)
Test cohort												
AdaBoost	92.59%	96.99%	94.99%	96.23%	94.01%	0.981(0.979-0.983)	82.92%	87.92%	85.46%	86.93%	84.17%	0.921(0.917-0.925)
XGBoost	92.40%	97.51%	95.18%	96.86%	93.90%	0.982(0.980-0.984)	83.95%	88.32%	86.17%	87.43%	85.04%	0.926(0.922-0.930)
DT	89.30%	97.10%	93.55%	96.26%	91.59%	0.959(0.956-0.962)	80.27%	87.84%	84.11%	86.47%	82.15%	0.896(0.891-0.901)
LR	90.00%	97.74%	94.22%	97.06%	92.14%	0.978(0.976-0.980)	79.43%	88.52%	84.04%	86.99%	81.64%	0.902(0.898-0.906)
RF	92.59%	97.28%	95.14%	96.59%	94.03%	0.982(0.980-0.984)	84.63%	87.56%	86.12%	86.82%	85.48%	0.924(0.920-0.928)
Validation cohort												
AdaBoost	89.48%	96.82%	94.78%	91.59%	95.98%	0.967(0.949-0.985)	83.52%	78.90%	80.45%	66.67%	90.46%	0.879(0.846-0.912)
XGBoost	89.84%	96.92%	94.95%	91.85%	96.11%	0.966(0.948-0.984)	85.65%	78.27%	80.75%	66.58%	91.52%	0.881(0.848-0.914)
DT	87.05%	96.44%	93.82%	90.48%	95.07%	0.947(0.925-0.969)	82.90%	81.81%	82.17%	69.76%	90.45%	0.875(0.841-0.909)
LR	89.33%	96.80%	94.72%	91.51%	95.92%	0.959(0.939-0.979)	83.83%	77.96%	79.93%	65.77%	90.52%	0.870(0.836-0.904)
RF	89.95%	96.52%	94.69%	90.89%	96.14%	0.963(0.944-0.982)	85.18%	76.81%	79.62%	64.99%	91.12%	0.874(0.840-0.908)

Abbreviations: SEN, sensitivity; SPE, specificity; ACC, accuracy; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve.

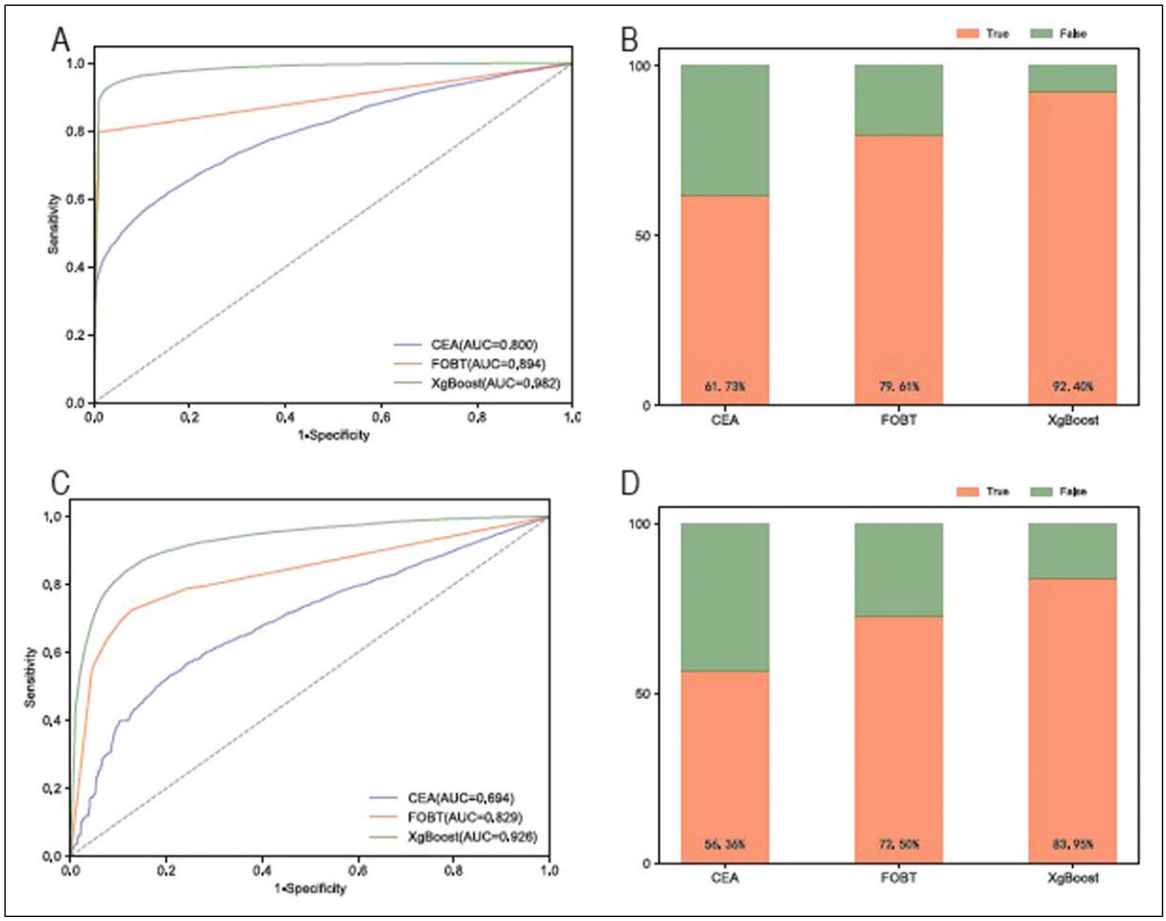

Figure 2. Performance Comparison of XGBoost, CEA, and FOBT in the Diagnosis of CRC. (A) ROC Curve Comparison of XGBoost, CEA, and FOBT in the Differential Diagnosis of CRC From HC; (B) True Positive Rates for XGBoost, CEA, and FOBT for all Patients With CRC Versus HC; (C) ROC Curve Comparison of XGBoost, CEA, and FOBT in the Differential Diagnosis of Polyp From HC; (D) True Positive Rates for XGBoost, CEA, and FOBT for all Patients With Polyps Versus HC; Abbreviations: XGBoost, Extreme Gradient Boosting; CEA, Carcinoembryonic Antigen; FOBT, Fecal Occult Blood Test; CRC, Colorectal Cancer; ROC, Receiver Operating Characteristic Curve; HC, Healthy Control; AUC, Area Under the Curve

77.55%, respectively. However, in the validation dataset, the XGBoost model showed the highest accuracy, precision, and F1 scores of 73.10%, 73.31%, and 71.56%, respectively. Meanwhile, all five algorithms showed excellent performance in the binary classification models (Table 1). Among HC vs CRC models, AdaBoost and XGBoost showed comparable performance, achieving accuracies of 94.78% and 94.95%, and AUCs of 0.967 and 0.966, respectively, within the validation dataset. In contrast, among Polys vs CRC models, the DT model achieved the highest accuracy of 82.17%, while the XGBoost model had the highest AUC of 0.881.

ML Models Exhibit Improved Diagnostic Performance Compared to CEA or FOBT for CRC

The XGBoost classifier, which demonstrated optimal discrimination performance primarily based on its AUC and sensitivity in the test set, exhibited superior discrimination compared to individual FOBT and CEA. Specifically, the AUC was 0.982 for the XGBoost HC vs CRC model, compared to 0.894 and 0.800 for FOBT and CEA, respectively (Figure 2A). This XGBoost classifier also achieved a significantly higher true positive rate (TPR) of 92.4% compared with those of FOBT (79.61%) and CEA (61.73%) (Figure 2B). In addition, the XGBoost Polys vs CRC classifier also exhibited significantly better diagnostic performance than FOBT and CEA, reaching an AUC of 0.926 and a TPR of 83.95%, respectively (Figure 2C, D).

Diagnostic Performance for CEA-Negative and FOBT-Negative CRC

Elevated CEA levels are often observed in the middle and late stages of CRC; therefore, patients showing high CEA may have missed the best treatment opportunity. For similar reasons, FOBT is not sensitive enough to detect early CRC.

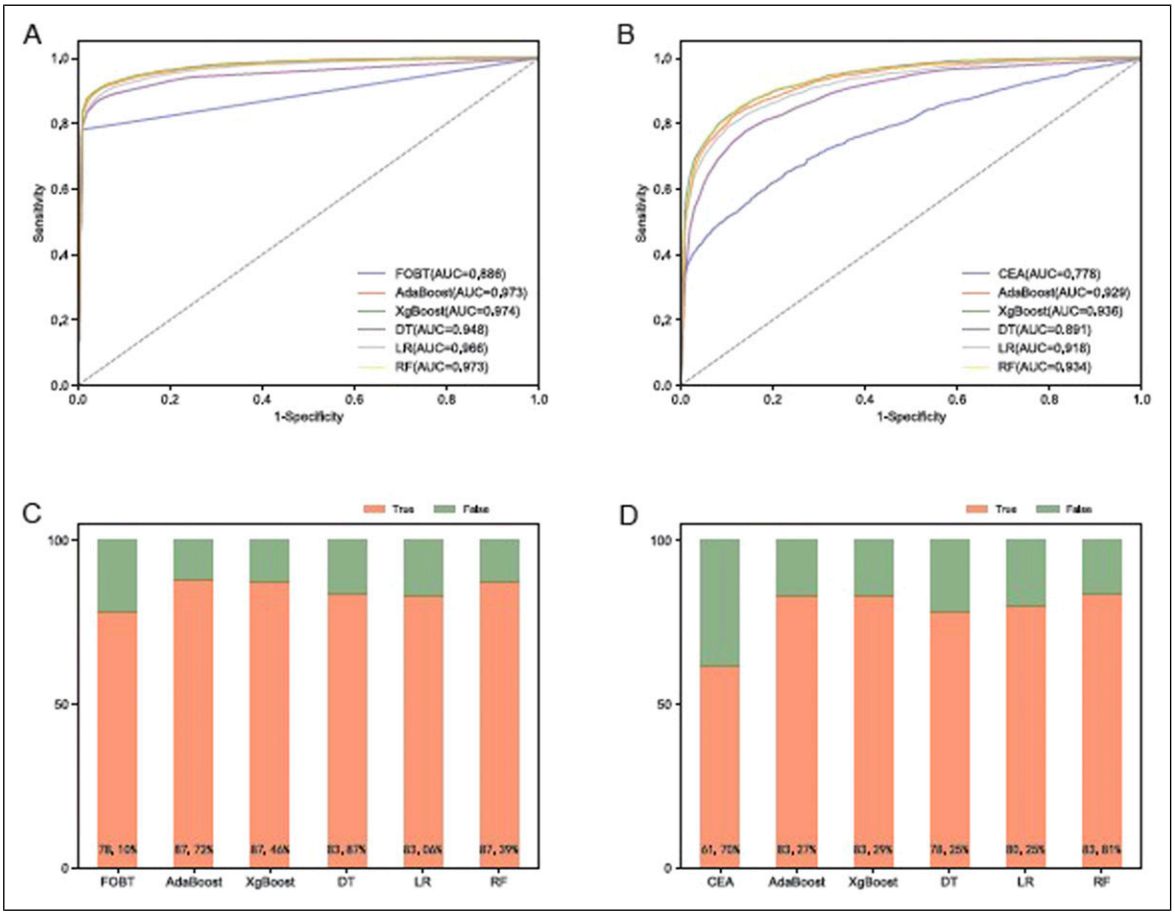

Figure 3. Performance of Five ML Models for CEA-Negative and FOBT-Negative CRC. (A) ROC of Five ML Models and FOBT for Patients With CEA-Negative CRC; (B) ROC of Five ML Models and CEA for Patients With FOBT-Negative CRC. (C) True Positive Rates for Five ML Models and FOBT for Patients With CEA-Negative CRC. (D) True Positive Rates for Five ML Models and CEA for Patients With FOBT-Negative CRC. Abbreviations: ML, Machine Learning; CEA, Carcinoembryonic Antigen (CEA < 5.0 ng/mL is Considered to be CEA-Negative); FOBT, Fecal Occult Blood Test; CRC, Colorectal Cancer; ROC, Receiver Operating Characteristic Curve; AUC, Area Under the Curve

CEA < 5.0 ng/mL was considered CEA-negative, and FOBT test results are categorized as either negative or positive. Among the 9428 CRC patients in cohort 1, 5174 (54.88%) were CEA-negative and 1922 were (20.39%) FOBT-negative. For HCs and patients with CEA-negative CRC, five ML algorithms, AdaBoost, XGBoost, DT, LR, and RF, exhibited excellent discrimination ability, and the AUCs were 0.973, 0.974, 0.948, 0.966, and 0.973, respectively (Figure 3A), significantly higher than the 0.886 of the single FOBT. In addition, five models maintained a high TPR for detecting patients with CEA-negative CRC with TPRs of 87.72%, 87.46%, 83.87%, 83.06%, and 87.39%, higher than that for FOBT (78.10%) (Figure 3C). In contrast, these patients showed no abnormalities when tested for CEA alone. Regarding the differential diagnosis of patients with FOBT-negative CRC, the five models still showed excellent performance with AUCs of 0.929, 0.936, 0.891, 0.918, and 0.934, compared to 0.778 for single CEA detection (Figure 3B) and TPRs 83.27%, 83.29%, 78.25%, 80.25%, and 83.81% vs 61.70% (Figure 3D).

Incorporating miR-92a Detection Improves the Diagnostic Performance of ML Models

By comparing stool miR-92a levels among HCs, patients with polyps, and those with CRC, we found that miR-92a levels were significantly higher in CRC patients than in both polyp patients and HCs (2.45 ± 1.01 vs 1.89 ± 0.74 and 1.54 ± 0.69, all $P < .0001$) (Figure 4A). Additionally, we observed a significant decrease in stool miR-92a levels following tumor resection in 50 CRC patients (3.01 ± 1.01 vs 1.76 ± 0.82, $P < .0001$) (Figure 4B), indicating that fecal miR-92a may be closely linked to the presence of CRC. The diagnostic specificity, sensitivity, and area under the curve (AUC) for miR-92a were 83.40%, 66.84%, and 0.783, respectively

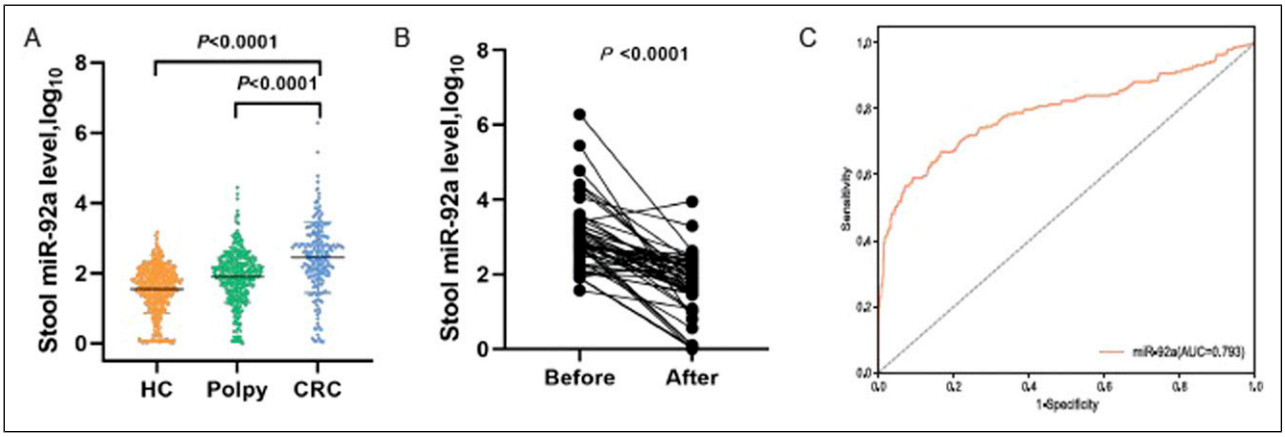

Figure 4. Expression Level of Fecal miR-92a. (A) Comparison of Levels of miR-92a in Stool Samples From Patients With CRC, Patients With Polyp, and HC; (B) Changes in Fecal miR-92a After Levels after Tumor Removal in Patients With CRC. *P* Values indicate Significant Differences Determined by the Wilcoxon Matched Pairs Test. (C) ROC Curve of miR-92a in the Differential Diagnosis of CRC From HC. Abbreviations: CRC, Colorectal Cancer; HC, Healthy Control; ROC, Receiver Operating Characteristic Curve

(Figure 4C). We then incorporated fecal miR-92a level to the 10 previously identified prediction factors and developed a new XGBoost model using the data from cohort 2. This addition improved the CRC diagnostic performance of the model; the AUC for distinguishing HCs from patients with CRC and polyps increased from 0.966 to 0.971 and from 0.865 to 0.919, respectively (Table 2).

Variable Importance and Interpretation

Based on the XGBoost CRC vs HC model, we visualized the effect of predictor variables on outcomes based on SHAP plots. The ranks of feature importance are shown in Figure 5A. FOBT, CEA, LYMPH%, and HCT were the most important features contributing to the risk of CRC. The specific contribution of each feature on the outcome is shown in Figure 5B, with blue and red representing negative and positive impact, respectively.

Development of a Prediction Program

A calculator program was constructed based on the 10 indicators using Python3.8, facilitating individualized prediction of prognostic risk in patients with CRC (Supplemental Figure S3). The source code can be downloaded at https://github.com/bushibenxin/calculation-tool.

Discussion

In this study, we employed five ML algorithms to develop classification models for diagnosis based on routine laboratory test results. The field of machine learning is replete with a multitude of algorithms, each possessing its own unique strengths and limitations. AdaBoost excels at managing high-dimensional data and large-scale features, and is relatively insensitive to outliers. However, it performs poorly in the presence of noise and outliers. XGBoost, a variant of the gradient boosting algorithm, is distinguished by its efficiency and scalability. Decision trees can process non-linear data effectively without the need for linear relationships between features, but they are susceptible to overfitting, particularly when the tree depth is excessive. Random forests mitigate the risk of overfitting by combining multiple decision trees, albeit at the expense of interpretability; individual decision trees are straightforward to interpret, but the overall interpretability of a random forest is diminished. Logistic regression is extensively

Table 2. ML Model Diagnostic Performance After Adding miR-92a for CRC

	HC vs CRC			Polyp vs CRC		
Model	SEN	SPE	AUC	SEN	SPE	AUC
XGBoost	89.84%	96.92%	0.966 (0.948-0.984)	67.15%	90.16%	0.865 (0.839-0.891)
XGBoost + miR-92a	86.15%	98.16%	0.971 (0.954-0.988)	80.96%	87.20%	0.919 (0.899-0.939)

Abbreviations: SEN, sensitivity; SPE, specificity; AUC, area under curves.

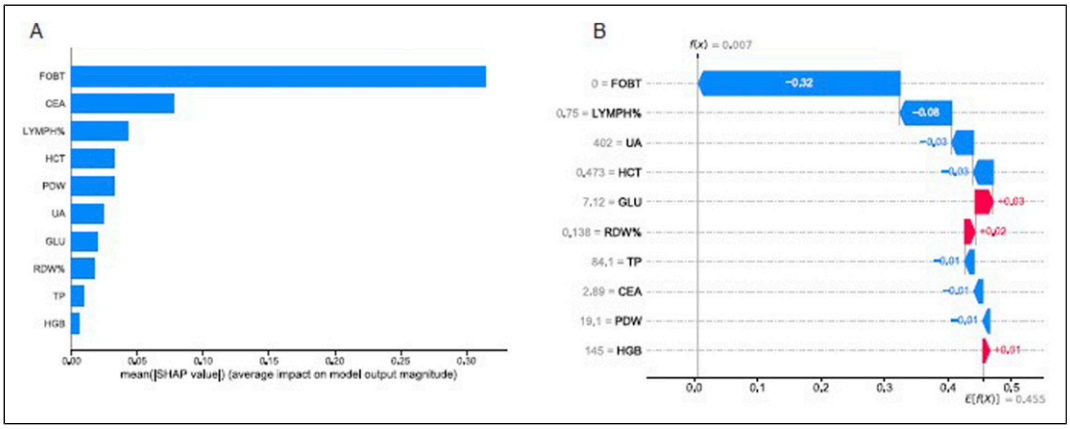

Figure 5. (A) Feature importance Derived From the XGBoost Model. (B) SHAP Analyses of the XGBoost Model for Predicting the Risk of CRC. Abbreviation: XGBoost, Extreme Gradient Boosting; SHAP, SHapley Additive exPlanations; CRC, Colorectal Cancer

used for binary classification problems, such as spam detection and disease prediction, but it may not yield optimal results for complex non-linear problems. Given the varying pros and cons of each algorithm, this study should utilize multiple algorithms for predicting colon cancer to ensure the precision of the prediction outcomes.

Specifically, we highlight that among the five ML models evaluated, XGBoost demonstrated the highest diagnostic accuracy, achieving an AUC of 0.966 for HC vs CRC and 0.881 for Polyp vs CRC, outperforming traditional biomarkers such as CEA and FOBT. Furthermore, XGBoost was able to identify CRC cases that tested negative for CEA or FOBT, demonstrating its potential clinical utility. The integration of stool miR-92a further improved the model's predictive power. Feature importance analysis using SHAP revealed that FOBT, CEA, LYMPH%, and HCT were the most influential variables in CRC risk prediction. Finally, we demonstrate that incorporating miR-92a, a novel CRC biomarker, may further enhance the diagnostic accuracy of the ML models. These findings suggest that ML-based approaches, particularly XGBoost, offer a more effective alternative to conventional screening methods and may facilitate early CRC detection, improving patient outcomes.

ML has been widely used in recent clinical studies to enhance diagnosis, predict prognosis, and support clinical decision-making. While many ML models concentrate on analyzing clinical images and identifying patterns—such as detecting tumors or lesions in MRI or endoscopic images—our study adopts a different approach by leveraging ML to extract insights from routine test results. These tests are more readily available in most clinical settings compared to advanced methods like imaging or genetic analyses. Our rationale for this approach includes two key points: first, there may be complex CRC-related patterns within routine data that only reveal clinical significance when analyzed in combination, which ML can effectively uncover; second, routine clinical tests are more accessible and cost-effective than advanced techniques, especially in underdeveloped regions. Moreover, their widespread use generates large datasets for training ML models, potentially enhancing diagnostic accuracy. Indeed, our ML models demonstrated excellent diagnostic performance compared to similar studies,[32-36] as displayed in Table 3. Although our aim is not to replace the current gold standard for CRC diagnosis—colonoscopy—our diagnostic models may help identify potential CRC patients during primary care or large-scale screenings. This would facilitate earlier referrals for colonoscopy, ultimately reducing the overall medical burden and improving early detection rates. Additionally, we have developed a computational tool for predicting CRC risk based on the optimal model was developed, designed for researchers with programming experience, which we anticipate will improve access to our ML models.

Despite the advantages of ML models, they are often criticized for their lack of explainability. The complexity of these models may obscure the factors contributing to the final classification results and the mechanisms behind them. To address this concern, our study employed clinically relevant features that were pre-selected using three different methods, thereby enhancing transparency and interpretability in the model's decision-making process. As anticipated, traditional CRC biomarkers such as CEA and FOBT were identified as key features and ranked highest in the SHAP analysis, reinforcing the robustness and clinical relevance of the ML model. Notably, variables related to nutritional deficiency—such as HGB, RDW%, TP, and HCT—and immune status (LYMPH%) also emerged as significant CRC-associated features.[37-39] Additionally, our research indicates that measuring GLU level aids in early diagnosis.[40] UA levels may be associated with an increased incidence of CRC, possibly because high serum UA is related to enhanced systemic inflammation and oxidative stress,[41,42] both of which are involved in CRC pathogenesis. However, it is important to note that none of these markers is specific to CRC, as various diseases and conditions may influence their levels. These findings further underscore the advantages of utilizing ML to

Table 3. Summary of This Study and Other Similar Research Findings

Reference	Population	Groups	Feature	Feature extraction method	Variable for prediction	Model	Best model and performance	External verification	Application program
			Source						
Our study	46 573 valid records: Healthy controls (HC) 11 793; Colorectal cancer (CRC) 9621; Polyp 10 125; Gastric cancer (GC) 10 714; Esophagus cancer (GC) 4319; China	HC; CRC; polyp; GC; EC	1415 testing indicators in clinical laboratories	Recursive feature elimination (RFE), Spearman correlation coefficients, and mutual information (MI)	LYMPH%; HGB; HCT; RDW%; PDW; TP; GLU; CEA; FOB; UA	Adaptive boosting (AdaBoost) Extreme gradient boosting (XGBoost); Decision tree (DT); Logistic regression (LR); Random forest (RF).	XGBoost model; for HC and CRC: sensitivity 89.84%, specificity 96.92%, and AUC 0.966; for CRC and polyp: sensitivity 85.65%, specificity 78.27%, and AUC 0.881; for GC and CRC, sensitivity 82.07%, specificity 81.97%, and AUC 0.835; for EC and CRC, sensitivity 88.39%, specificity 71.30%, and AUC 0.910	Yes	Yes
Li et al[28]	1164 electronic medical records (CRC patients: 582; healthy controls: 582); China	Healthy controls; CRC patients	Laboratory data, including liver enzymes, lipid profiles, complete blood counts, and tumor biomarkers	Spearman and LR	CEA; HGB; Lp (a); HDL	LR; RF; k-nearest neighbors (KNN), support vector machine (SVM); Naïve Bayes (NB)	LR model; ; AUC: 0.865, sensitivity: 89.5%, specificity: 83.5%, PPV: 84.4%, NPV: 88.9%	No	No
Long et al[29]	340 CRC patients and 134 paired noncancerous Controls; Gene expression Omnibus (GEO)	Non-cancerous controls; CRC	Microarray-based gene expression	Area under the curve (AUC)-RF, Boruta, and Vita	39, 41, and 68 individual gene markers from AUCRF, Boruta, and Vita, respectively.	RF; LR: NB; KNN	RF model; Mean accuracy 0.998 (standard deviation (SD) < 0.003), mean specificity 0.999 (SD < 0.003), and mean sensitivity 0.998 (SD < 0.004)	Yes	No
Nakajima et al[30]	201 CRCs and 31 non-CRCs; Tokyo	CRCs; controls; Benign	Seven kinds of polyamines	AUC	Urinary polyamines	AD tree	Combinations of polyamines AUC value of 0.961 (95% CI: 0.937-0.984, $P < 0.0001$)	NO	No

(continued)

Table 3. (continued)

Reference	Population	Groups	Feature	Feature extraction method	Variable for prediction	Model	Best model and performance	External verification	Application program
			Source						
Kinar et al[31]	606 403 Israelis (of whom 3135 were diagnosed with CRC) and a case-control UK dataset of 5061 CRC cases and 25 613 controls; Israeli and UK	Cancer-free; CRC	Blood counts	Linear regression model	20 blood count parameters	Gradient boosting Model; RF	The specificity was 88 ± 2% in the Israeli validation set and 94 ± 1% in the UK dataset. Detecting 50% of CRC cases, the odds ratio was 26±5 and 40±6, respectively, for a false-positive rate of 0.5%. Specificity for 50% detection was 87 ± 2% a year before diagnosis and 85 ± 2% for localized cancers	Yes	No
Zhao et al[32]	89 healthy individuals and 92 CRC patients. France and Germany data from the NCBI website	Healthy controls; CRC patients	Location, age, and gender, and BMI and tumor type, tumor grade, and DNA	LR; AUC	Firmicutes; Bacteroidetes; BMI; age	SVM	RBF kernel type; accuracy 90.1% when k = 5, and 91.2% when k = 10	No	No

uncover critical yet hidden patterns and clinical indicators for cancer diagnosis. The differential indicators we have identified in the real world may also provide a scientific research foundation for subsequent research on the CRC mechanism. The selection of these parameters elucidates the model's explainability from the perspective of model input. In terms of model output, we also compared the predictive performance of the constructed model with traditional indicators. It is noteworthy that the predictive AUC and TRP of the model constructed in this study are both higher than those of traditional indicators. Furthermore, the five models we constructed offer multiple indicators, including sensitivity, specificity, accuracy, and AUC, facilitating user understanding and selection based on individual needs. From both the input and output perspectives, our model exhibits a high degree of explainability.

The diagnostic accuracy of ML models can be further improved by incorporating additional biomarkers. Our introduction of miR-92a resulted in enhanced diagnostic accuracy for these models. Although stool miR-92a is not part of routine analysis, it has shown a strong correlation with the presence of CRC and its treatment. Research indicates that miR-92a plays a critical role in the invasion and migration of CRC, suggesting its potential involvement in CRC development.[43,44] It is important to note that, in the early stages of CRC, epigenetic modifications are more likely to occur than gene mutations; therefore, miR-92a may serve as a promising next-generation diagnostic biomarker for colonic polyps and malignant tumors.[45] Currently, miR-92a monitoring has been implemented in clinical laboratories across some countries and regions. This real-world application holds significant importance for our study. For patients who have undergone miR-92a testing, incorporating the test results into our prediction models can markedly enhance the models' predictive accuracy.

Several limitations of this study should be acknowledged. First, while the selected diagnostic indicators are reliable, caution is warranted when generalizing the findings to populations outside of China or to different age groups. Second, the study did not distinguish between adenomatous and serrated polyps. Although serrated colorectal polyps were traditionally considered benign, they are now recognized as precursors to approximately one-third of CRC cases. Further research is needed to develop predictive models that account for the different types of polyps.[46] Third, since the predictive variables were obtained retrospectively, there is a risk of data leakage into the fitted models, which could affect the evaluation. These results should be validated in a prospective study to ensure their robustness.

Future studies should concentrate on refining biomarker signatures through the integration of proteomic, genomic, and metabolomic analyses. Kaplan-Meier and survival analysis can validate the model's clinical significance by assessing long-term outcomes and stratifying patients based on prognostic factors. Additionally, external validation across diverse populations, cost-effectiveness analysis, and integration into clinical decision support systems will ensure the model's reliability, economic feasibility, and practical usability in healthcare settings, ultimately improving early CRC detection and patient management.

Conclusion

In this study, we employed five ML algorithms to develop classification models for diagnosis based on routine laboratory test results. Although these models were constructed using only the ten most important indicators, they achieved impressive accuracies in distinguishing healthy controls from CRC patients and in differentiating polyp patients from CRC patients in both the test and validation datasets, with the XGBoost model exhibiting the highest accuracy. Notably, the ML models outperformed traditional diagnostic markers such as CEA and FOBT. Additionally, the ML models were able to identify CRC patients who tested negative for CEA and FOBT, highlighting their potential for early CRC diagnosis. Finally, we demonstrate that incorporating miR-92a, a novel CRC biomarker, may further enhance the diagnostic accuracy of the ML models. In the future, by exploring the integration of multi-omics and employing deep learning methods, we aim to further enhance prediction accuracy and clinical applicability.

Acknowledgements

The opinions on this document are those from the authors alone and do not represent the views of the institutions to which they belong.

ORCID iD

Liu Yang https://orcid.org/0009-0000-1859-5397

Ethical Approval

This study was approved by the Medical Ethics Committee of the First Affiliated Hospital of the Air Force Medical University of the People's Liberation Army, located in Xi'an, China, on April 12, 2024, with the approval number KY20242119 and was conducted in compliance with the Guidelines for Good Clinical Practice and the Declaration of Helsinki. Informed consent was waived due to the study's retrospective design. The studies were conducted in accordance with the local legislation and institutional requirements.

Author Contributors

RL: Literature search, Data collection, Data analysis, Model construction, Writing –original draft. XH: Literature search, Data collection, Data analysis, Software, Model construction. LY: Designed, Supervision, Writing – review & editing. JL: Funding acquisition, Project administration, Writing – review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This work was

supported by the Innovation Capability Support Program of Shaanxi (Program No. 2021LXZX3-01).

Declaration of Conflicting Interests

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

Data Availability Statement

Data are available upon reasonable request. The data presented in this study are available upon request from the corresponding author.

Supplemental Material

Supplemental material for this article is available online.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72:7-33.
2. Krishnan ST, Winkler D, Creek D, et al. Staging of colorectal cancer using lipid biomarkers and machine learning. *Metabolomics*. 2023;19:84.
3. Alnabulsi A, Murray GI. Proteomics for early detection of colorectal cancer: recent updates. *Expert Rev Proteomics*. 2018; 15:55-63.
4. Han YD, Oh TJ, Chung TH, et al. Early detection of colorectal cancer based on presence of methylated syndecan-2 (SDC2) in stool DNA. *Clin Epigenet*. 2019;11:51.
5. Maida M, Macaluso FS, Ianiro G, et al. Screening of colorectal cancer: present and future. *Expert Rev Anticancer Ther*. 2017; 17:1131-1146.
6. Parekh PJ, Shams R, Oldfield EC, Nicholas JJ, Johnson DA. Computed tomography colonography: ready for prime time for colon cancer screening? *J Clin Gastroenterol*. 2014;48:745-751.
7. Zygulska AL, Pierzchalski P. Novel diagnostic biomarkers in colorectal cancer. *Int J Mol Sci*. 2022;23:852.
8. Bretthauer M, Wieszczy P, Løberg M, et al. Estimated lifetime gained with cancer screening tests: a meta-analysis of randomized clinical trials. *JAMA Intern Med*. 2023;183: 1196-1203.
9. Campos-da-Paz M, Dórea JG, Galdino AS, Lacava ZGM, de Fatima Menezes Almeida Santos M. Carcinoembryonic antigen (CEA) and hepatic metastasis in colorectal cancer: update on biomarker for clinical and biotechnological approaches. *Recent Pat Biotechnol*. 2018;12:269-279.
10. Raza A, Khan AQ, Inchakalody VP, et al. Dynamic liquid biopsy components as predictive and prognostic biomarkers in colorectal cancer. *J Exp Clin Cancer Res*. 2022;41:99.
11. Jayasinghe M, Prathiraja O, Caldera D, et al. Colon cancer screening methods: 2023 update. *Cureus*. 2023;15:e37509.
12. Knudsen AB, Rutter CM, Peterse EFP, et al. Colorectal cancer screening: an updated modeling study for the US preventive services task Force. *JAMA*. 2021;325:1998-2011.
13. Yang Y, Meng WJ, Wang ZQ. MicroRNAs in colon and rectal cancer - novel biomarkers from diagnosis to therapy. *Endocr, Metab Immune Disord: Drug Targets*. 2020;20:1211-1226.
14. Sun K, Han R, Han Y, Shi X, Hu J, Lu B. Accuracy of combined computed tomography colonography and dual energy iiodine map imaging for detecting colorectal masses using high-pitch dual-source CT. *Sci Rep*. 2018;8:3790.
15. Malla M, Loree JM, Kasi PM, Parikh AR. Using Circulating Tumor DNA in Colorectal Cancer: Current and Evolving Practices. *J Clin Oncol*. 2022;40:2846-2857.
16. Chen F, Dai X, Zhou CC, et al. Integrated analysis of the faecal metagenome and serum metabolome reveals the role of gut microbiome-associated metabolites in the detection of colorectal cancer and adenoma. *Gut*. 2022;71:1315-1325.
17. Nannini G, Meoni G, Tenori L, et al. Fecal metabolomic profiles: a comparative study of patients with colorectal cancer vs adenomatous polyps. *World J Gastroenterol*. 2021;27:6430-6441.
18. Tan B, Qiu Y, Zou X, et al. Metabonomics identifies serum metabolite markers of colorectal cancer. *J Proteome Res*. 2013; 12:3000-3009.
19. Huang D, Sun W, Zhou Y, et al. Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer Metastasis Rev*. 2018;37:173-187.
20. Huang A, Sun Z, Hong H, et al. Novel hypoxia- and lactate metabolism-related molecular subtyping and prognostic signature for colorectal cancer. *J Transl Med*. 2024;22:587.
21. Schmitt M, Greten FR. The inflammatory pathogenesis of colorectal cancer. *Nat Rev Immunol*. 2021;21:653-667.
22. Zhao R, Xia D, Chen Y, et al. Improved diagnosis of colorectal cancer using combined biomarkers including Fusobacterium nucleatum, fecal occult blood, transferrin, CEA, CA19-9, gender, and age. *Cancer Med*. 2023;12:14636-14645.
23. Deo RC. Machine learning in medicine. *Circulation*. 2015;132: 1920-1930.
24. Peiffer-Smadja N, Rawson TM, Ahmad R, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect*. 2020;26:584-595.
25. Du Y, McNestry C, Wei L, Antoniadi AM, McAuliffe FM, Mooney C. Machine learning-based clinical decision support systems for pregnancy care: a systematic review. *Int J Med Inf*. 2023;173:105040.
26. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*. 2023;186:1772-1791.
27. Rodó X, San-José A, Kirchgatter K, López L. Changing climate and the COVID-19 pandemic: more than just heads or tails. *Nat Med*. 2021;27:576-579.
28. Feng R, Liu X, Chen J, Chen DZ, Gao H, Wu J. A deep learning approach for colonoscopy pathology WSI analysis: accurate segmentation and classification. *IEEE J Biomed Health Inform*. 2021;25:3700-3708.
29. Prezja F, Annala L, Kiiskinen S, et al. Improving performance in colorectal cancer histology decomposition using deep and ensemble machine learning. *Heliyon*. 2024;10:e37561.
30. Choi HH, Cho YS, Choi JH, Kim HK, Kim SS, Chae HS. Stool-based miR-92a and miR-144* as noninvasive biomarkers for colorectal cancer screening. *Oncology*. 2019;97:173-179.

31. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370:1453-1457.
32. 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.
33. 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;20:296.
34. 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;19:756.
35. 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;23:879-890.
36. Zhao D, Liu H, Zheng Y, He Y, Lu D, Lyu C. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Med Biol Eng Comput*. 2019;57:901-912.
37. Alsalman A, Al-Mterin MA, Abu-Dayeh A, Alloush F, Murshed K, Elkord E. Associations of complete blood count parameters with disease-free survival in right- and left-sided colorectal cancer patients. *J Pers Med*. 2022;12:816.
38. Bossi P, Delrio P, Mascheroni A, Zanetti M. The spectrum of Malnutrition/Cachexia/Sarcopenia in oncology according to different cancer types and settings: a narrative review. *Nutrients*. 2021;13:1980.
39. 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.
40. Schoen RE, Tangen CM, Kuller LH, et al. Increased blood glucose and insulin, body size, and incident colorectal cancer. *J Natl Cancer Inst*. 1999;91:1147-1154.
41. Kimura Y, Tsukui D, Kono H. Uric acid in inflammation and the pathogenesis of atherosclerosis. *Int J Mol Sci*. 2021;22:12394.
42. Gherghina ME, Peride I, Tiglis M, Neagu TP, Niculae A, Checherita IA. Uric acid and oxidative stress-relationship with cardiovascular, metabolic, and renal impairment. *Int J Mol Sci*. 2022;23:3188.
43. Wei QD, Zheng WB, Sun K, Xue Q, Yang CZ, Li GX. MiR-92a promotes the invasion and migration of colorectal cancer by targeting RECK. *Int J Clin Exp Pathol*. 2019;12:1565-1577.
44. Shi Y, Liu Z. Serum miR-92a-1 is a novel diagnostic biomarker for colorectal cancer. *J Cell Mol Med*. 2020;24:8363-8367.
45. Yau TO, Tang CM, Harriss EK, Dickins B, Polytarchou C. Faecal microRNAs as a non-invasive tool in the diagnosis of colonic adenomas and colorectal cancer: a meta-analysis. *Sci Rep*. 2019;9:9491.
46. Carballal S, Balaguer F, Jeg IJ. Serrated polyposis syndrome; epidemiology and management. *Best Pract Res Clin Gastroenterol*. 2022;58-59:101791.

Appendix

Abbreviation

CRC	colorectal cancer
ML	machine learning
HC	healthy control
Polyp	polyp patient
AdaBoost	adaptive boosting
XGBoost	extreme gradient boosting
DT	decision tree
LR	logistic regression
RF	random forest
CEA	carcino-embryonic antigen
FOBT	fecal occult blood testing
SHAP	shapley additive explanations
LYMPH%	lymphocyte percentage
HCT	hematocrit
ctDNA	circulating tumor DNA
miRNA	microRNA
FIT-DNA	fecal multi-target DNA
UPLC-MS	ultra-performance liquid chromatography-mass spectrometry
NMR	nuclear magnetic resonance spectroscopy
GC-TOFMS	gas chromatography time-of-flight mass spectrometry
cfDNA	cell-free DNA
FDA	Food and Drug Administration
CV	computer vision
CNN	convolutional neural network
LOINC	Logical Observation Identifiers Names and Codes
KNN	K-nearest neighbor
RFE	recursive feature elimination
MI	mutual information
AUC	area under the curve
ROC	receiver operating characteristic
HGB	hemoglobin
RDW%	red blood cell distribution width
PDW	platelet distribution width
TP	total protein
GLU	glucose
UA	uric acid
TPR	true positive rate