DOI: 10.66069/ojspub.27450739

Machine Learning for Predicting Colorectal Precancerous Lesions: A Comparative Study

Shuxin Tong[1], Jiong Wu[1], Xuerong Zhang[1], Jinpeng You[1], Haijuan Xiao[2,3,*]

[1] Shaanxi University of Chinese Medicine, Xianyang 712046, China
[2]Integrated Traditional Chinese and Western Medicine Hospital of Wenzhou Medical University, Wenzhou 325000, China
[3]Oncology Hospital, Affiliated Hospital of Shaanxi University of Chinese Medicine, Xianyang 712000, China
**Correspondence Author*

Abstract: ***Early screening and identification of precancerous lesions in colorectal cancer (CRC) are crucial for improving prognosis, and the application of machine learning in this field has achieved significant progress. This article systematically reviews the multimodal features of CRC precancerous lesions, covering multiple dimensions such as endoscopic images, molecular markers, genomics, gut microbiota, and Traditional Chinese Medicine (TCM) syndrome patterns. It further compares the performance and applicability of three types of prediction methods: traditional machine learning, deep learning, and fusion learning. On this basis, the article proposes that by constructing standardized multimodal datasets, developing data fusion and causal inference methods, and enhancing model interpretability, it is possible to improve the accuracy and clinical translation potential of prediction models, providing theoretical and technical references for early warning and precise prevention of CRC.***

Keywords: Colorectal Precancerous Conditions, Machine Learning, Prediction Model, Early Diagnosis and Treatment.

1. Introduction

Colorectal cancer (CRC) is a common gastrointestinal malignancy in China. According to statistics, in 2022, there were approximately 517,000 new cases of CRC and about 240,000 deaths in China. Its incidence and mortality rates ranked second and fourth nationwide, respectively, presenting a serious challenge for prevention and control [1,2]. Between 2003 and 2015, the five-year survival rate for CRC patients in China was at most 56.9%, significantly lower than that in countries such as Europe, the United States, Japan, and South Korea. However, early detection and timely intervention can increase the five-year survival rate to over 90% [3]. Therefore, the early and accurate identification of and intervention in CRC precancerous lesions is key to reducing the incidence and mortality of CRC.

The vast majority of CRC cases are sporadic, with more than 90% following the "adenoma-to-cancer" pathway, progressing from colorectal adenomas (CRA), which are the most common precancerous lesions in the general population; serrated lesions are another important class of parallel precancerous lesions [3]. It is currently widely accepted that CRC follows a multistage model of "normal mucosa → precancerous lesions (adenomas/serrated lesions) → invasive carcinoma," a process driven by a combination of genetic mutations, epigenetic alterations, imbalances in the gut microbiome, and lifestyle factors such as diet and obesity [4]. The "Chinese Expert Consensus on Early Diagnosis and Treatment of Colorectal Cancer (2023 Edition)" also classifies Inflammatory Bowel Disease (IBD) as a high-risk factor for CRC; individuals in this group require more intensive endoscopic surveillance.

In recent years, researchers have employed multidisciplinary approaches—including endoscopy, pathology, molecular biology, and epidemiology—to study early biomarkers of colorectal cancer (CRC) precancerous lesions, providing a basis for early warning and precision intervention in CRC [3]. Against this backdrop, with the rapid development of big data and artificial intelligence technologies, computational methods centered on machine learning have been extensively explored in the fields of early CRC screening, lesion identification, and risk prediction [5,6]. However, the predictive performance of these models is often constrained by factors such as data modality and quality, algorithm selection and optimization, and the size and representativeness of the datasets, thereby limiting their generalization ability and potential for clinical translation [7]. Therefore, this paper will begin with a multimodal feature analysis to systematically elucidate the key characteristics of CRC precancerous lesions, compare the performance and applicable scenarios of different machine learning methods, summarize current challenges, and outline future development directions, thereby providing a reference for the construction of an efficient and reliable AI-assisted decision-making system.

2. Multimodal Analysis of Precancerous Lesions in CRC

Progressive adenomas and serrated lesions with malignant potential are key precancerous stages of colorectal cancer (CRC) [8,9]; their early identification and resection are central to CRC prevention and control, which currently relies primarily on colonoscopy [8]. A systematic understanding of the multimodal characteristics of CRC precancerous lesions—encompassing clinical epidemiology, endoscopic pathology, molecular microbiology, and Traditional Chinese Medicine (TCM) constitution—can provide a basis for variable selection and algorithm choice in predictive models.

2.1 Clinical and Epidemiological Characteristics

The development of CRC is associated with various risk factors, including genetic background, lifestyle, age, gender, and socioeconomic status. Genetics and family history constitute a clear predisposing factor; approximately 2–8% of cases are associated with genetic syndromes such as Lynch syndrome, and having a first-degree relative with the disease

can double an individual's risk [10]. Patients with IBD also have a significantly increased risk of cancer due to long-term inflammation of the intestines [11]. Among lifestyle factors, high intake of red and processed meats, a low-fiber diet, obesity, physical inactivity, smoking, and alcohol consumption have all been shown to promote carcinogenesis [12]. Age is a strong risk factor; approximately 90% of cases occur in individuals over 50 years of age, but incidence rates among younger populations are rising in some countries, suggesting that screening strategies need to be optimized [13]. Gender differences are also pronounced: men have a higher incidence and poorer prognosis, while women are more prone to developing right-sided colon cancer [12]. Furthermore, socioeconomic factors further influence the distribution of the disease through complex interactions with the aforementioned risk factors, reflecting social inequalities in areas such as dietary patterns and access to healthcare [14,15].

2.2 Endoscopic and Histopathological Features

CRC precancerous lesions present with diverse endoscopic appearances. Colorectal adenomas typically appear as polypoid elevations; pathological types include tubular, villous, or mixed. Among these, villous adenomas and those with larger diameters (>2 cm) carry a higher risk of malignant transformation, and the risk of malignant transformation is significantly elevated in cases with histological dysplasia [16]. Among serrated lesions, sessile serrated adenomas (SSA/P) are most commonly found in the right half of the colon; they have a pale surface, indistinct borders, and are covered by a mucus cap. Traditional serrated adenomas (TSA) predominantly occur in the left half of the colon and the rectum, presenting as elevated polyps with characteristic serrated structures visible on histological examination; When either type is accompanied by dysplasia, the risk of malignancy rises sharply [17,18]. IBD-associated dysplasia is classified into flat and elevated types; it is often associated with long-term intestinal mucosal inflammation, and the incidence of CRC is 5 to 10 times higher than in the general population [19]. These features collectively form an important data foundation for constructing risk prediction models.

2.3 Biological Characteristics

2.3.1 Genomic and Epigenetic Characteristics

In the classic "adenoma-to-cancer sequence," inactivation of the tumor suppressor gene APC leads to sustained activation of the Wnt/β-catenin signaling pathway, thereby initiating the disease process [20]; KRAS mutations cause uncontrolled cell proliferation, accelerating tumor formation [21]; and TP53 inactivation further drives the progression of CRC precancerous lesions to cancer [22]. In the "serrated pathway," the BRAF V600E mutation is characterized by an accompanying extensive CpG island methylation phenotype (CIMP), which often silences mismatch repair genes (such as MLH1), leading to microsatellite instability (MSI) and thereby accelerating genomic mutations [23–25], driving the progression of colorectal precancerous lesions to cancer. Furthermore, epigenetic alterations, such as dysregulation of histone modifications, are prevalent in both pathways; they jointly regulate gene expression, creating conditions conducive to tumor initiation and progression [26,27].

2.3.2 Gut Microbiota and Microenvironment Characteristics

In the development of CRC precancerous lesions, an imbalance in the intestinal microbiome primarily manifests as a disruption in microbial community structure—specifically, a decrease in the number of beneficial bacteria, an increase in the abundance of harmful bacteria, and a reduction in overall microbial diversity [28]. Fusobacterium nucleatum, a harmful bacterium closely associated with CRC, is found at significantly elevated levels in precancerous lesions and patient feces. It promotes local inflammation and cell proliferation by activating signaling pathways such as NF-κB and Wnt; toxins secreted by enterotoxin-producing Bacteroides fragilis can disrupt the intestinal epithelial barrier, trigger intestinal inflammation, and drive the progression of precancerous lesions [29]. Furthermore, impaired intestinal barrier function allows bacteria and their metabolites to translocate into the bloodstream, further exacerbating inflammation and immune dysregulation [30]. At the metabolic level, a reduction in short-chain fatty acids (SCFAs) produced by the gut microbiota can weaken its anti-inflammatory and barrier-protective functions [31], while an increase in secondary bile acids promotes cell proliferation and inhibits apoptosis by activating receptors such as farnesoid X receptor (FXR), thereby collectively driving the progression of precancerous lesions [32].

2.3.3 Biochemical and Molecular Markers

In terms of fecal biomarkers, the Fecal Immunochemical Test (FIT) offers higher specificity than traditional chemical methods and is less susceptible to interference from diet and medications [33]; Fecal DNA testing, by analyzing mutations in genes such as KRAS and BRAF as well as the methylation status of genes such as SEPT9, not only enables non-invasive screening but can also detect lesions in the proximal colon and flat lesions that are easily missed by colonoscopy; however, its specificity is lower [34]. Among blood biomarkers, carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA19-9) may be elevated even in some patients with precancerous lesions; in addition to being used for disease monitoring and prognosis assessment, they also provide a reference for early auxiliary diagnosis [35]. Furthermore, novel liquid biopsy markers such as circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) can dynamically reflect the molecular characteristics of tumors and show promise in early detection and monitoring. However, due to limitations such as high costs and a lack of standardization, they are currently used primarily for research or in specific clinical settings [36].

2.4 Characteristics of Traditional Chinese Medicine

2.4.1 Disease Names, Etiology, and Pathogenesis in Traditional Chinese Medicine

Although Traditional Chinese Medicine does not have specific terms for "colorectal adenomas" or "serrated lesions," based on their clinical characteristics, they can be classified under categories such as "intestinal mass," "accumulation," and "polyps." Traditional Chinese Medicine holds that precancerous lesions of CRC are often caused by exposure to

damp pathogenic factors, irregular diet, and emotional imbalances, which lead to weakness of the spleen and stomach. This impairs their ability to transport and transform nutrients, resulting in the internal generation of damp turbidity that obstructs the intestines and viscera. Over time, this leads to qi stagnation and blood stasis; the intermingling of dampness and stasis forms polyps, and in severe cases, this stagnation may transform into toxins and evolve into cancer. "The Ling Shu Shi" states: "Dregs and filth accumulated in the intestines coalesce to form polyps, just as dampness and heat evaporating and stagnating can cause fungi to grow on wood. This is why it is called 'intestinal accumulation'." This demonstrates that dampness stagnating in the intestines and spleen deficiency with dampness stagnation are the core pathomechanisms, with the primary site of the disease being the large intestine and close associations with the spleen and liver [37].

2.4.2 Characteristics of Syndromes in Traditional Chinese Medicine

The distribution of syndrome patterns in CRC precancerous lesions varies by age: among middle-aged and young adults, "excess" patterns or a mixture of "excess" and "deficiency" predominate, with common patterns including accumulation of damp-heat, obstruction by cold-dampness, and qi stagnation with blood stasis; among the elderly, "deficiency" patterns or a mixture of "deficiency" and "excess" predominate, with common patterns including spleen qi deficiency and spleen deficiency with dampness and blood stasis [38]. For patients diagnosed via colonoscopy but without obvious symptoms, TCM can conduct risk assessments based on constitutional differentiation. Common predisposing constitutions include yang deficiency, qi deficiency, phlegm-dampness, and damp-heat [37]. Treatment emphasizes addressing both the root cause and the symptoms, focusing primarily on tonifying qi and strengthening the spleen, supplemented by methods such as clearing heat and draining dampness, warming and transforming cold-dampness, and promoting qi circulation and blood activation to eliminate pathogenic factors. Because dampness and blood stasis are deeply entrenched and difficult to eliminate, this condition is prone to recurrence; therefore, strengthening the spleen to eliminate dampness and promoting blood circulation to resolve stasis are also important therapeutic approaches for preventing recurrence [39].

Based on the foregoing analysis of multimodal features, it can be observed that the assessment of CRC precancerous lesions relies on data derived from multiple sources, with each feature type exhibiting inherent limitations with regard to detection cost, invasiveness, and accuracy. Therefore, integrating multimodal data constitutes a key strategy for constructing efficient and cost-effective machine learning models and enabling early warning for CRC. A detailed comparison of the multimodal features of CRC precancerous lesions is presented in Table 1.

Table 1: Comparison of Multimodal Characteristics of CRC Precursor Lesions

Feature dimension	Characteristic	Advantage	Disadvantage
Endoscopy - Pathological Morphology	Observe the mucosal morphology and lesion structure intuitively; Pathological biopsy to clarify the type of lesion, degree of dysplasia, etc	The ability to directly visualize lesions is the gold standard for diagnosis	Invasive operation with risks of perforation and bleeding; Operation relies on physician experience; Not suitable for large-scale screening
Biochemical and molecular markers	Detect specific substances in blood and feces (FIT, CEA, CA19-9, fecal DNA, ctDNA, etc.)	Non invasive, easy to operate, low cost, suitable for large-scale initial screening	The positive rate of some biomarkers in the precancerous stage is low; Insufficient specificity in fecal DNA testing; Easy to be disturbed by diet and medication
Genome and Epigenetics	Detect mutations and methylation of APC, KRAS, BRAF, MSI, etc	Revealing lesion risk at the molecular level, clarifying genetic susceptibility background, and applicable for early warning	The detection cost is high, the technology is complex, and the clinical popularization is limited
Gut microbiota and microenvironment	Analyze the structure, metabolites, and intestinal barrier function of beneficial bacteria	Non invasive, reflecting the imbalance of intestinal microbiota	Inconsistent testing standards; Easy to be influenced by diet and medication
Epidemiology	Analyze risk factors such as genetic background, lifestyle, age, gender, etc	Revealing the distribution characteristics of diseases and providing a basis for the formulation of population screening strategies	Due to the influence of sample representativeness and survey methods, it cannot be directly used for accurate individual risk assessment
Traditional Chinese Medicine Constitution and Syndrome	Analyzing the etiology and pathogenesis from the perspectives of physical constitution and syndrome types	Emphasize the overall state and individual differences, complementing the characteristics of modern medicine	Lack of objective detection indicators; Diagnosis relies on physician experience

3. Comparison of Models for Predicting CRC Precursor Lesions Using Machine Learning Methods

Machine learning is a core branch of artificial intelligence that aims to enable computers to learn autonomously and optimize their performance through data-driven approaches [40]. Its mainstream methods can be divided into two categories: traditional machine learning and deep learning. Each category has its own distinct characteristics in terms of feature processing, performance, and applicable scenarios [41].

3.1 Traditional Machine Learning Methods for Predicting CRC Precancerous Lesions

Traditional machine learning relies on manual feature engineering and algorithm selection, and excels at analyzing structured data such as clinical indicators and demographic information. Commonly used algorithms include logistic regression (LR), support vector machines (SVM), random forests (RF), and gradient-boosted decision trees (XGBoost).

Traditional machine learning has demonstrated consistent performance in predicting the risk of CRC precancerous lesions. For example, one study used random forests (RF) to integrate multiple indicators—including age, sex, BMI, and

history of diabetes—to construct a colon polyp prediction model, achieving an AUC of 0.79 in external validation and thus providing a viable tool for initial screening in asymptomatic populations [42]. Another study, involving over 3,000 patients, systematically compared various algorithms and found that the XGBoost-based model offered the best predictive performance (AUC = 0.780). The same study also revealed significant population heterogeneity: subgroup models performed particularly well for individuals under 60 years of age and for men (with AUCs of 0.788 and 0.793, respectively), and key predictive indicators varied across different subgroups [43]. These findings suggest that traditional machine learning can achieve more precise risk stratification through subgroup optimization.

Furthermore, in the prediction of pedunculated serrated lesions in the colorectal tract, one study used LASSO regression to identify 14 key indicators—including age and hemoglobin—and constructed a logistic regression (LR) model with an AUC of 0.79; SHAP analysis was used to identify the core predictive factors, thereby enhancing clinical reliability [44]. Overall, traditional machine learning is suitable for small- to medium-scale structured data and offers advantages such as high computational efficiency and model transparency; however, feature extraction relies on expert experience, and its ability to process unstructured data is limited.

3.2 Deep Learning Methods for Predicting CRC Precancerous Lesions

Deep learning methods use multi-layer neural networks to perform end-to-end automatic feature extraction and are particularly effective at processing high-dimensional, unstructured data such as endoscopic images and histological slides.

In colonoscopy image analysis, one study used a CNN model with 43 convolutional layers to distinguish between normal intestinal mucosa, adenomatous polyps, and adenocarcinoma, achieving an accuracy of 94.39% [45]. Another study, based on architectures such as VGG19 and ResNet101, achieved 100% accuracy in distinguishing between colorectal adenocarcinoma and benign tissue [46]. Furthermore, to reduce reliance on expert-labeled data, GAN T et al. [47] proposed the FPSiam self-supervised learning framework. By pretraining the model on large-scale unlabeled colonoscopy videos, the model learns general visual feature representations, significantly reducing the reliance on labeled data for downstream polyp detection while demonstrating superior performance.

In addition to image-based tasks, deep learning has also been extended to clinical text analysis. Transformer-based models have been used to automatically extract information related to CRC precancerous lesions from electronic health records, thereby aiding in the construction of high-quality clinical databases [48]. In summary, deep learning performs exceptionally well in both image and text tasks; however, it relies on large-scale annotated data, incurs high computational costs, and the “black-box” nature of the models reduces clinical interpretability.

3.3 Ensemble Learning Methods for Predicting CRC Precancerous Lesions

A single data source or model often struggles to fully capture the complex mechanisms underlying the onset and progression of CRC. Therefore, integrating multimodal data and combining multiple models for decision-making has become a key approach to improving predictive performance.

Multimodal data fusion aims to integrate complementary, heterogeneous information sources and can effectively overcome the limitations of single-modality data. Depending on the stage at which fusion occurs, it can be classified into types such as feature-level and decision-level fusion [49]. One study developed a microbial model based on gut microbiota biomarkers for the diagnosis of CRC and CRA, which demonstrated robust performance across multi-regional populations (CRC AUC: 0.817–0.867; CRA AUC: 0.766–0.833). After further integration of clinical features, the AUC of the integrated model increased to 0.939 and 0.925, respectively, significantly outperforming the single-model approaches [50]. Furthermore, contemporary research is attempting to elucidate TCM theories at the molecular level. For example, some studies have linked the TCM therapeutic principle of “dispelling dampness and clearing toxins” to the regulation of the LHPP/PI3K/AKT signaling pathway and the modulation of the gut microbiota, providing a biological basis for the development of predictive models that integrate Chinese and Western medicine [51].

Table 2: Comparison of the Characteristics of CRC Precursor Lesions Predicted Using Three Types of Machine Learning Methods

Method type	Core characteristics	Advantage	Disadvantage	Applicable scenarios
Traditional machine learning methods	Dependent on artificial feature engineering, suitable for feature association mining of low dimensional structured data	Simple model, fast training speed, low computational cost, and strong interpretability	Feature extraction relies on expert knowledge and is difficult to handle high-dimensional unstructured data such as images/texts	Risk prediction and initial screening based on clinical indicators and epidemiological characteristics
Deep learning methods	End to end feature learning, suitable for complex nonlinear pattern mining of high-dimensional unstructured data such as images/texts	Automatically extract complex features with high recognition accuracy	Relying on experts to annotate data, high computational costs, and poor interpretability	Endoscopic imaging/pathological image analysis, medical text processing
Fusion learning method	Multi modal/multi model integration, capturing complex multidimensional information correlations	Improve model robustness and overall performance	Further reduce complexity and model interpretability	Multi modal data integration and auxiliary diagnosis of complex scenarios

Multi-model ensemble approaches enhance generalization capabilities by combining the strengths of different algorithms. For example, in predicting rectal cancer tumor deposition, a voting model that integrates RF, XGBoost, and other algorithms achieved an AUC of 0.875, demonstrating the potential of ensemble strategies in complex tasks [52]. In summary, by synergistically leveraging heterogeneous data and diverse models, fusion learning methods hold promise for overcoming the biases inherent in single information sources and more comprehensively capturing the complex evolutionary trajectories from precancerous lesions to early-stage cancer. This represents a cutting-edge direction for advancing the precise risk assessment of CRC precancerous lesions.

4. Methods for Improving the Accuracy of Predicting Precancerous Lesions Using CRC

Currently, machine learning-based predictive models still face challenges in the identification of CRC precancerous lesions, such as inconsistent data quality, limited modalities, unclear causality, and insufficient interpretability. To enhance model performance and clinical utility, future research could focus on the following four aspects:

4.1 Constructing High-quality Standardized Multimodal Datasets

Current research predominantly relies on single-center retrospective data, which has limited sample sizes and selection bias, making it difficult to fully reflect population heterogeneity. Additionally, multimodal data such as endoscopic images, molecular biomarkers, and TCM syndromes lack unified standards, constraining model generalizability. Future efforts should focus on advancing multicenter prospective cohort studies, establishing standardized protocols for clinical, imaging, molecular, and TCM data collection and annotation, with particular emphasis on developing objective TCM-related data acquisition tools. By leveraging privacy-preserving technologies like federated learning, cross-institutional collaboration can be achieved while ensuring data security, thereby providing a large-scale, high-quality dataset foundation for model training [53,54].

4.2 Developing Deep Fusion Methods for Multimodal Data

The occurrence of CRC precancerous lesions involves multidimensional factors, making it difficult for a single data source to comprehensively capture their complex mechanisms. Existing models primarily rely on simple feature concatenation, failing to fully leverage the complementary value of multi-source information. Future research should actively explore deep integration technologies for multimodal data, such as multi-source feature alignment based on attention mechanisms, cross-modal representation learning, and graph neural network fusion, to achieve efficient integration of multidimensional data including endoscopic images, genomics, gut microbiota, and traditional Chinese medicine constitution [55]. By employing strategies like joint training and multi-task learning, models can enhance their understanding and utilization of heterogeneous data, thereby more comprehensively and accurately identifying precancerous lesion risks.

4.3 Exploring Causal Inference Frameworks Based on Real-world Data

Most existing models only reveal the “correlation” between features and risks, rather than “causality,” which limits their intervention guidance value [56]. Real-world data is often accompanied by confounding factors, missing values, and selection bias, which limits the application of traditional causal inference methods. Therefore, it is necessary to introduce a causal learning framework that adapts to the characteristics of medical data, such as combining double difference, propensity score matching, structural causal models, etc., to infer the causal effects of key risk factors from observational data [56,57]. By constructing a causal oriented prediction model, not only can prediction stability be improved, but it can also provide mechanistic basis for early intervention, promoting the evolution of CRC prevention and control from “correlation warning” to “causal intervention”.

4.4 Enhancing Model Interpretability and Clinical Trustworthiness

Although complex models such as deep learning have superior performance, they exhibit a “black box” characteristic, which reduces the trust of clinical doctors. Future research should focus on model interpretability design and output transparency, integrate explanatory tools such as LIME and SHAP, and visualize key predictive features and their contributions [58]. Especially when integrating subjective data such as traditional Chinese medicine, clear decision-making criteria and pathological explanations need to be provided. By developing a human-machine collaborative interaction system, clinical doctors can participate in model iteration and validation, enhance their sense of identification with prediction results, and promote the reliable integration and widespread application of artificial intelligence assisted systems in clinical practice [59].

References

[1] HAN B, ZHENG R, ZENG H, et al. Cancer incidence and mortality in China, 2022[J]. J Natl Cancer Cent. 2024, 4(1):47-53.

[2] BRAY F, LAVERSANNE M, SUNG H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: A Cancer Journal for Clinicians, 2024, 74(3):229-263.

[3] Expert Group on Early Diagnosis and Treatment of Cancer, Chinese Society of Oncology, Chinese Medical Association. Expert consensus on the early diagnosis and treatment of colorectal cancer in China (2023 edition) [J]. National Medical Journal of China, 2023, 103(48): 3896-3908.

[4] Liu J H, Wang R P, Zhu C Z, et al. Research progress of precancerous lesions of colorectal tumors: insights and breakthroughs [J]. Chinese Journal of Coloproctology, 2025(8): 68-74.

[5] SCHÖLER J, ALAVANJA M, DE LANGE T, et al. Impact of AI-aided colonoscopy in clinical practice: A

prospective randomised controlled trial[J]. BMJ open gastroenterology, 2024, 11(1): e001247.
[6] Miao Y X, Wang X C, Wang H X, et al. Application of artificial intelligence-assisted systems in the endoscopic screening of colorectal cancer[J]. Journal of Multidisciplinary Cancer Management (Electronic Version), 2025, 11(1):10-20.
[7] MOHAMMAD E B, AHMAD M. A systematic evaluation of big data-driven colorectal cancer studies[J]. Medicinski Glasnik: Official Publication of the Medical Association of Zenica-Doboj Canton, Bosnia and Herzegovina, 2024, 21(1): 63-77.
[8] Wang H Y. Research progress of advanced colorectal adenoma[J]. Advances in Clinical Medicine, 2024, 14(1): 292-296.
[9] SAWICKI T, RUSZKOWSKA M, DANIELEWICZ A, et al. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis[J]. Cancers, 2021, 13(9): 2025.
[10] VALLE L, VILAR E, TAVTIGIAN S V, et al. Genetic predisposition to colorectal cancer: Syndromes, genes, classification of genetic variants and implications for precision medicine. [J]. The Journal of pathology, 2019, 247(5): 574-588.
[11] HNATYSZYN A, HRYHOROWICZ S, KACZMAREK-RYŚ M, et al. Colorectal carcinoma in the course of inflammatory bowel diseases[J]. Hereditary Cancer in Clinical Practice, 2019, 17: 18.
[12] Chen J, Xiang Y B. Progress on risk factors for early-onset colorectal cancer: a review of prospective epidemiological studies[J]. Chinese Journal of Cancer Prevention and Treatment, 2025(6).
[13] STOFFEL E M, MURPHY C C. Epidemiology and mechanisms of the increasing incidence of colon and rectal cancers in young adults[J]. Gastroenterology, 2020, 158(2): 341-353.
[14] CARETHERS J M, DOUBENI C A. Causes of socioeconomic disparities in colorectal cancer and intervention framework and strategies[J]. Gastroenterology, 2020, 158(2): 354-367.
[15] DENNY L, JEMAL A, SCHUBAUER-BERIGAN M, et al. Social inequalities in cancer risk factors and health-care access[M]//Reducing social inequalities in cancer: evidence and priorities for research. International Agency for Research on Cancer, 2019.
[16] CORLEY D A, JENSEN C D, MARKS A R, et al. Adenoma detection rate and risk of colorectal cancer and death[J]. The New England Journal of Medicine, 2014, 370(14): 1298-1306.
[17] YIN A, ZENG Z, WAN X, et al. Jejunal serrated adenoma diagnosed and treated by double-balloon enteroscopy[J]. Case Reports in Gastroenterology, 2018, 12(2): 528-531.
[18] CHEN B, SCURRAH C R, MCKINLEY E T, et al. Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps[J]. Cell, 2021, 184(26): 6262-6280.e26.
[19] Ordás I, Eckmann L, Talamini M, et al. Ulcerative colitis[J]. Lancet, 2012, 380(9853):1606-1619.
[20] BIAN J, DANNAPPEL M, WAN C, et al. Transcriptional regulation of wnt/β-catenin pathway in colorectal cancer[J]. Cells, 2020, 9(9): 2125.
[21] NAJUMUDEEN A K, FEY S K, MILLETT L M, et al. KRAS allelic imbalance drives tumour initiation yet suppresses metastasis in colorectal cancer in vivo[J]. Nature Communications, 2024, 15(1): 100.
[22] BROCKMUELLER A, BUHRMANN C, MORAVEJOLAHKAMI A R, et al. Resveratrol and p53: How are they involved in CRC plasticity and apoptosis? [J]. Journal of Advanced Research, 2024, 66: 181-195.
[23] GROTHEY A, FAKIH M, TABERNERO J. Management of BRAF-mutant metastatic colorectal cancer: A review of treatment options and evidence-based guidelines[J]. Annals of Oncology: Official Journal of the European Society for Medical Oncology, 2021, 32(8): 959-967.
[24] MÜLLER D, GYŐRFFY B. DNA methylation-based diagnostic, prognostic, and predictive biomarkers in colorectal cancer[J]. Biochimica Et Biophysica Acta. Reviews on Cancer, 2022, 1877(3): 188722.
[25] HELDERMAN N C, ANDINI K D, VAN LEERDAM M E, et al. MLH1 promotor hypermethylation in colorectal and endometrial carcinomas from patients with lynch syndrome[J]. The Journal of Molecular Diagnostics: JMD, 2024, 26(2): 106-114.
[26] ASHKTORAB H, BELGRAVE K, HOSSEINKHAH F, et al. Global histone H4 acetylation and HDAC2 expression in colon adenoma and carcinoma[J]. Digestive Diseases and Sciences, 2009, 54(10): 2109-2117.
[27] GEZER U, USTEK D, YÖRÜKER E E, et al. Characterization of H3K9me3- and H4K20me3-associated circulating nucleosomal DNA by high-throughput sequencing in colorectal cancer[J]. Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine, 2013, 34(1): 329-336.
[28] GARVEY M. Intestinal dysbiosis: Microbial imbalance impacts on colorectal cancer initiation, progression and disease mitigation[J]. Biomedicines, 2024, 12(4): 740.
[29] CAI J, SUN L, GONZALEZ F J. Gut microbiota-derived bile acids in intestinal immunity, inflammation, and tumorigenesis[J]. Cell Host & Microbe, 2022, 30(3): 289-300.
[30] SUZUKI T. Regulation of the intestinal barrier by nutrients: The role of tight junctions[J]. Animal Science Journal = Nihon Chikusan Gakkaiho, 2020, 91(1): e13357.
[31] FUSCO W, LORENZO M B, CINTONI M, et al. Short-chain fatty-acid-producing bacteria: Key components of the human gut microbiota[J]. Nutrients, 2023, 15(9): 2211.
[32] NIE Q, LUO X, WANG K, et al. Gut symbionts alleviate MASH through a secondary bile acid biosynthetic pathway[J]. Cell, 2024, 187(11): 2717-2734.e33.
[33] MEKLIN J, SYRJÄNEN K, ESKELINEN M. Colorectal cancer screening with traditional and new-generation fecal immunochemical tests: A critical review of fecal occult blood tests[J]. Anticancer Research, 2020, 40(2): 575-581.
[34] IMPERIALE T F, PORTER K, ZELLA J, et al. Next-generation multitarget stool DNA test for colorectal cancer screening[J]. The New England Journal of Medicine, 2024, 390(11): 984-993.

[35] Xu L Y, Jiang N, Zhao J B, et al. Value of serum HSP90α combined with lncRNA SNHG5 and conventional serum tumor markers in the diagnosis of early colorectal cancer[J]. Hainan Medical Journal, 2024, 35(19): 2792-2797.
[36] DENG Z, WU S, WANG Y, et al. Circulating tumor cell isolation for cancer diagnosis and prognosis[J]. EBioMedicine, 2022, 83: 104237.
[37] Zhang X L, Hui Y F, Fan M M, et al. Progress in the Prevention and Treatment of Colorectal Adenoma by Traditional Chinese Medicine[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2022, 24(2): 490-494.
[38] Chen L R, Chen Y, Huang S, et al. Pathogenic Factors, Characteristics and Distribution of TCM Syndrome Analysis of 450 Patients with Colorectal Polyps[J]. Liaoning Journal of Traditional Chinese Medicine, 2018, 45(1): 9-12, insert page 1.
[39] Ye H Z, Cheng Y, Zhong C L, et al. Exploration of the Thoughts for the Prevention and Treatment of Postoperative Recurrence of Colorectal Adenoma Based on the Pathogenesis of Latent Toxin[J]. Journal of Guangzhou University of Traditional Chinese Medicine, 2023, 40(6): 1527-1531.
[40] KÜHL N, GOUTIER M, HIRT R, et al. Machine learning in artificial intelligence: Towards a common understanding[J]. Logical Techniques in computer science, 2024, 2(2): 53-59.
[41] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[A]. arXiv, 2015.
[42] AVRAM M F, LUPA N, KOUKOULAS D, et al. Random forests algorithm using basic medical data for predicting the presence of colonic polyps[J]. Frontiers in Surgery, 2025, 12.
[43] Zhang Y. Study on risk prediction model of colorectal polyps based on machine learning[D]. Chongqing: Chongqing Medical University, 2024.
[44] Huang S H, Chen G X. Construction of an interpretable risk prediction model for colorectal sessile serrated lesion based on machine learning algorithms[J]. Journal of Xuzhou Medical University, 2025, 45(10): 753-759.
[45] PARK H C, KIM Y J, LEE S W. Adenocarcinoma recognition in endoscopy images using optimized convolutional neural networks[J]. Applied Sciences, 2020, 10(5): 1650.
[46] IRAWATI I D, ANDREA LARASATY I, HADIYOSO S. Comparison of convolution neural network architecture for colon cancer classification[J]. International Journal of Online and Biomedical Engineering (iJOE), 2022, 18(3): 164-172.
[47] GAN T, JIN Z, YU L, et al. Self-supervised representation learning using feature pyramid siamese networks for colorectal polyp detection[J]. Scientific Reports, 2023, 13(1): 21655.
[48] WIEST I C, WOLF F, LESSMANN M E, et al. LLM-AIx: An open source pipeline for information extraction from unstructured medical text based on privacy preserving large language models[J]. medRxiv: the Preprint Server for HEALTH Sciences, 2024: 2024.9.2.24312917.
[49] SUN X, TIAN Y, LU W, et al. From single- to multi-modal remote sensing imagery interpretation: A survey and taxonomy[J]. Science China Information Sciences, 2023, 66(4): 140301.
[50] Zhou D, Chen Y, Wang Z, et al. Integrating clinical and cross-cohort metagenomic features: a stable and non-invasive colorectal cancer and adenoma diagnostic model[J]. Front Mol Biosci. 2024 Jan 22;10:1298679.
[51] Guo C, Liu X, Xu Y, et al. Exploring the Mechanism of Action of Canmei Formula Against Colorectal Adenoma Through Multi-Omics Technique[J]. Front Cell Dev Biol. 2021 Nov 29;9:778826.
[52] Wang J, Hu F, Li J, et al. Comparative performance of multiple ensemble learning models for preoperative prediction of tumor deposits in rectal cancer based on MR imaging[J]. Sci Rep. 2025 Feb 10;15(1):4848.
[53] Sun X D, Lu C X, Gao T, Xie C T, Tang Z T, Zhao Y Y. Privacy computing under cross agency collaboration — Feasibility analysis of federated learning [J]. Chinese Journal of Computer Application, 2024, 40(8): 111-113.
[54] OYEKAN B. Developing privacy-preserving federated learning models for collaborative health data analysis across multiple institutions without compromising data security[J]. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2024, 3(3): 139-164.
[55] Waqas A, Tripathi A, Ramachandran RP, et al. Multimodal data integration for oncology in the era of deep neural networks: a review[J]. Front Artif Intell. 2024 Jul 25;7:1408843.
[56] ROSENBAUM P R, RUBIN D B. The central role of the propensity score in observational studies for causal effects[J]. Biometrika, 1983, 70(1): 41-55.
[57] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology [J]. Epidemiology. 2000 Sep; 11(5):550-60.
[58] Sathyan A, Weinberg AI, Cohen K. Interpretable AI for bio-medical applications[J]. Complex Eng Syst. 2022 Dec;2(4):18.
[59] Park, HJ., Kim, S.H., Choi, J.Y. et al. Human–machine cooperation meta-model for clinical diagnosis by adaptation to human expert's diagnostic characteristics [J]. Sci Rep 13, 16204 (2023).