Alqalam Journal of Medical and Applied Sciences. 2025;8(3):1555-1563
https://doi.org/10.54361/ajmas.258342

Original article

A Systematic Review of Machine Learning Techniques for The Diagnosis of Colorectal Cancer

Huweida Darbi*, Rabab Algadhy

Department of Computer, Faculty of Sciences, University of Derna, Derna, Libya
Corresponding Email. *huweida.darbi@uod.edu.ly*

Abstract
In recent years, have attended surpassing developments in the area of machine learning in various fields, particularly in the medical sector. The use of machine learning technologies has become a promising tool for supporting diagnosis, prediction, and clinical decision-making, contributing to improving the quality of healthcare and reducing human error. One of the important applications which have emerged is the use of machine learning technologies in cancer diagnosis, given the critical importance of this field in increasing survival rates and improving treatment results. Colorectal cancer is one of the most common and risky types of cancer worldwide, ranking top in terms of incidence and mortality rates. Early detection of colon cancer is a crucial factor in improving survival rates. This study focuses attention on the need for intelligent tools capable of supporting clinical decisions based on exact and in-depth analyses of medical data. Machine learning techniques have demonstrated a high ability to analyze the vast amount of clinical data, radiological images, and genetic patterns associated with colon cancer, opening up new possibilities for achieving more accurate early diagnosis compared to traditional methods. This paper presents a systematic survey of the available literature that uses machine learning techniques in colon cancer diagnosis, which will help identify innovations applied in this research area and explore future trends. This paper aims to conduct a literature survey of machine learning techniques in colon cancer diagnosis using the SLR methodology, analyze and compare the literature, and identify the appropriate technique to address issues in colon cancer diagnosis.

Keywords: Colon Cancer Diagnosis, Machine Learning (ML), Machine Learning Techniques.

Introduction

Colorectal cancer (CRC) is one of the most common and dangerous types of cancer in the world, affecting both men and women. It usually begins with abnormal growths, like adenomatous polyps, in the colon or rectum, which can become cancerous over time [1]. As stated by new global statistics, CRC is the third most commonly diagnosed cancer and the second leading cause of cancer-related deaths, with more than 1.9 million new cases and approximately 935,000 deaths stated in 2020 alone [2]. Traditionally, CRC is diagnosed through screening methods such as colonoscopy, fecal occult blood testing (FOBT), and sigmoidoscopy. Despite significant progress made by traditional diagnostic methods in detecting colon cancer, there remains a significant gap in access to accurate and effective diagnoses in the early stages of the disease, which is a critical factor in improving survival rates and reducing mortality rates. Because of this challenge, ML techniques have emerged as promising tools in the field of medical diagnostics, including the early detection of colorectal cancer, that can help doctors by analyzing complex data and detecting subtle patterns that may be difficult to identify using traditional methods.

Machine learning (ML) is a branch of artificial intelligence that purposes to develop systems able of learning from data and identify patterns without explicit programming for each task [3]. ML techniques are generally categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning [4]. These techniques are capable of examining huge quantities of clinical, biological, and genetic data to identify patterns associated with CRC development.

In the medical field, ML has become a critical tool for analyzing complex medical data such as radiological images, electronic health records, and genomic information [5][6]. Its main applications include accurate and fast disease diagnosis, predicting disease progression, clinical decision support, and the analysis of medical images such as X-rays and MRI scans [7][21]. Several studies have proved the effectiveness of ML models in enhancing diagnostic accuracy, predicting disease stages, and identifying critical risk factors [18][21]. Due to their capacity to learn from genetic data and make accurate predictions, ML techniques have become crucial to the advancement of predictive and precision medicine, displaying significant potential for improving early CRC detection and reducing the overall healthcare and economic burden [19].

Several studies have been conducted on machine learning techniques in colon cancer diagnosis using various algorithms. Accordingly, the research problem is to explore the extent to which machine learning techniques can improve the accuracy of colon cancer diagnosis and which techniques are best for colon cancer diagnosis. The main prompting for this research stems from the urgent need to explore the true potential of machine learning techniques in improving the diagnostic process and to determine the extent to which these techniques can reduce reliance on individual physician expertise and accelerate access to accurate diagnoses. In addition, machine learning and its techniques have become the subject of many new studies, enabling many investigators to conduct new research. However, figuring out the optimal technique

***Received**: 30-05-2025 - **Accepted**: 24-07-2025 - **Published**: 31-07-2025*

remains difficult. To achieve this, we need a systematic study that provides us with the latest research in this field.

Methods

A systematic literature review method is a well-defined approach to identifying, evaluating, and interpreting all relevant studies concerning a specific research question, topic area, or phenomenon of interest. Therefore, it is important to understand how to perform it efficiently and reliably [22]. This method was chosen because we wanted to obtain a fair, reliable, and unbiased evaluation of a particular method.

Research questions

We first formulated the review goal through Goal-Question Metric perspectives (purpose, issue, viewpoint).
Purpose: Analysis and characterization.
Issue: Identify machine learning techniques used to diagnose colon cancer.
Viewpoint: From the researcher's perspective.
Based on the purpose, we derived the following research questions:
RQ1: What are the published studies on machine learning techniques in colon cancer diagnosis between 2021 and 2025?
Rationale: This question aims to identify aspects of the literature related to machine learning techniques in colon cancer between 2021 and 2025.
RQ2: What are the best machine learning techniques used in colon cancer diagnosis?
Rationale: This question aims to explore the techniques used and identify the best machine learning methods from recent reports on colon cancer.
RQ3: What is the success rate and impact of machine learning on colon cancer diagnosis?
Rationale: This question focuses on identifying current research interests and concerns of researchers.
What are the challenges and future research directions identified in these studies?

Search process

The research methodology aims to identify and review prominent machine learning techniques used in colon cancer diagnosis. For this research, we selected four digital repositories to identify relevant articles: IEEE Xplore Digital Library, Google Scholar, ACM Digital Library, and Science Direct. The search query ensures that any research paper in the database meets basic quality criteria: originality, high impact, and a high h-index.

After formulating the research questions, it is necessary to create an effective search string to find relevant studies from the electronic databases used. This study included studies from 2021 to 2025 to focus on emerging methods. A general search string was created to target the developed research questions. To identify a large number of studies that use machine learning techniques in colon cancer diagnosis, specific search terms were selected. For the key terms, we have provided a list of synonyms, abbreviations, and alternative words. To create the search terms, the main keywords and their synonyms were linked using the "OR" and "AND" operators. Search terms: "Machine learning" AND "colon cancer" OR "colon cancer diagnosis using Machine Learning Techniques" OR "Machine Learning Techniques" OR "Colonoscopy". The given set of search terms was used to extract the desired results from the selected digital repositories. The selected places in the browsed journals are listed in (Table 1).

Table 1: Journals that were searched and their rankings

ID	Journals	h-index	Impact factor	SJR	Class
J1	Scientific Reports	347	4.13	1.24	Q1
J2	European Journal of Cancer (EJC)	193–256	7.6	2.50–2.69	Q1
J3	Sensors	273	3.4	0.764	Q1
J4	NPJ Precision Oncology	46	6.8	3.370	Q1
J5	BMC Cancer	171	3.4	1.178	Q2
J6	Journal of Electrical Systems	22	0.50	0.180	Q4
J7	PLOS ONE	467	2.6	0.803	Q1
J8	Computers in Biology and Medicine	142	6.3	1.447	Q1
J9	Biology: Journal of International Biological Sciences	129	4	2.715	Q1
J10	Journal of Informatics in Medicine Unlocked	66	4.21	0.762	Q2

Received*: 30-05-2025 -* ***Accepted****: 24-07-2025 -* ***Published****: 31-07-2025*

J11	Cancers: Journal of Oncology	157	4.8	1.462	Q1
J12	JMIR Cancer	28	2.7	1.025	Q2
J13	Cancer Control	83	2.6	0.881	Q2

Inclusion criteria and exclusion criteria

The selection of primary studies is based on the inclusion criteria (IC), while the exclusion criteria (EC) are used to exclude studies. The inclusion and exclusion criteria used during the review are listed in (Table 2), and (Figure 1) illustrates the general review workflow.

Quality criteria

Each primary study was assessed for quality using a checklist. We use a three-point scale to answer each question, either as "yes", "to some extent", or "no". By including "to some extent," we avoided neglecting statements where authors provided only limited information to answer the assessment questions. Each quality assessment question was answered by assigning a numerical value (1 = "yes", 0 = "no", and 0.5 = "to some extent"). The quality assessment questions are defined as:

Q1: Is the study objective clearly defined?
Q2: Is the machine learning technique used clearly described? (e.g., Random Forest, SVM, CNN, etc.)
Q3: Is the type of data used clearly defined? (e.g., medical images, laboratory data, demographic data, etc.)
Q4: Is the sample size sufficient to support the results?
Q5: Are the performance metrics reported? (e.g., accuracy, sensitivity, specificity, area under the curve)
Q6: Is the study published in a reliable scientific source (a peer-reviewed journal, a recognized scientific conference)?

Table 2: Inclusion and Exclusion Criteria.

Inclusion Criteria		**Exclusion Criteria**	
IC1	The study must be published in the English language and full-text article.	**EC1**	The studies that are not available on the selected electronic databases.
IC2	The studies that focus on Machine learning techniques used in colon cancer diagnosis.	**EC2**	The duplicated articles of the same study.
IC3	The studies were published between 2021 and 2025 time periods.	**EC3**	The studies lack in answering the devised research questions.
IC4	The academic publications should correctly be "peer-reviewed journals" or "academic conferences"	**EC4**	Magazine papers, short papers, poster papers, editorials, tutorials,non-reviewed papers, editorials, and presentations are excluded.
IC5	The studies that cover the devised research questions.		

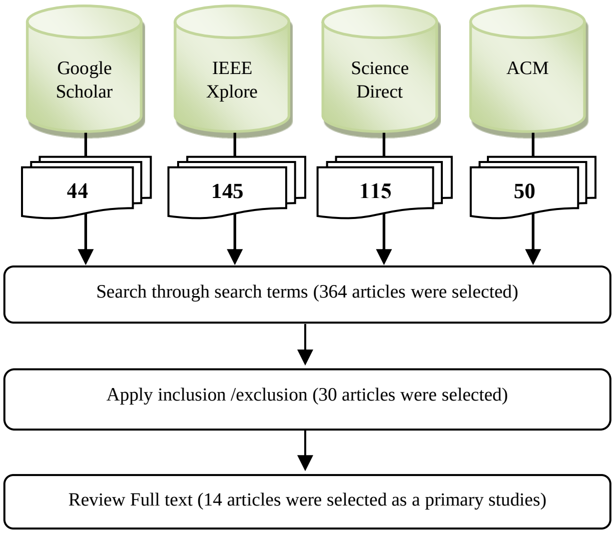
Google Scholar
IEEE Xplore
Science Direct
ACM
44
145
115
50
Search through search terms (364 articles were selected)
Apply inclusion /exclusion (30 articles were selected)
Review Full text (14 articles were selected as a primary studies)

Figure 1: Article selection process.

***Received**: 30-05-2025 - **Accepted**: 24-07-2025 - **Published**: 31-07-2025*

Conducting the review

Primary studies selection

A comprehensive search was conducted, and a set of articles was extracted from all the selected databases, as described in (Figure 1). After these steps and after excluding duplicated manuscripts and reviewing titles and abstracts, the filtration output reached 30. Finally, applying quality assessment, irrelevant papers were excluded, and as a result, we had 14 selected papers.

Table 3: Publications on colon cancer diagnosis using machine learning

Ref.	Title	Year	Type	Id-Type
[8]	Predicting Early-Onset Colorectal Cancer in Individuals Below Screening Age Using Machine Learning and Real-World Data: Case Control Study"	2025	Journal	J12
[9]	Colon cancer diagnosis by means of of explainable deep learning	2024	Journal	J1
[10]	A Machine Learning Approach for Detection and Classification of Colon Cancer using Convolutional Neural Network Architecture	2024	Journal	J6
[11]	Machine Learning as a Tool for Early Detection: A Focus on Late-Stage Colorectal Cancer across Socioeconomic Spectra	2024	Journal	J11
[12]	An interpretable machine learning system for colorectal cancer diagnosis from pathology slides	2024	Journal	J4
[13]	Explainable machine learning models for colorectal cancer prediction using clinical laboratory data	2025	Journal	J13
[14]	Colorectal Polyp Image Detection and Classification through Grayscale Images and Deep Learning	2021	Journal	J3
[15]	Machine learning-based colorectal cancer prediction using global dietary data	2023	Journal	J5
[16]	A machine learning tool for identifying non-metastatic colorectal cancer in primary care	2023	Journal	J2
[17]	Accurate prediction of colorectal cancer diagnosis using machine learning based on immunohistochemistry pathological images	2024	Journal	J1
[18]	Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis	2022	Journal	J8
[19]	Machine Learning-Based Identification of Colon Cancer Candidate Diagnostics Genes	2022	Journal	J9
[20]	Detection of effective genes in colon cancer: A machine learning approach	2021	Journal	J10
[21]	A practical approach for colorectal cancer diagnosis based on machine learning	2025	Journal	J7

Figure 2 shows the number of publications related to using Machine learning techniques in colon cancer.

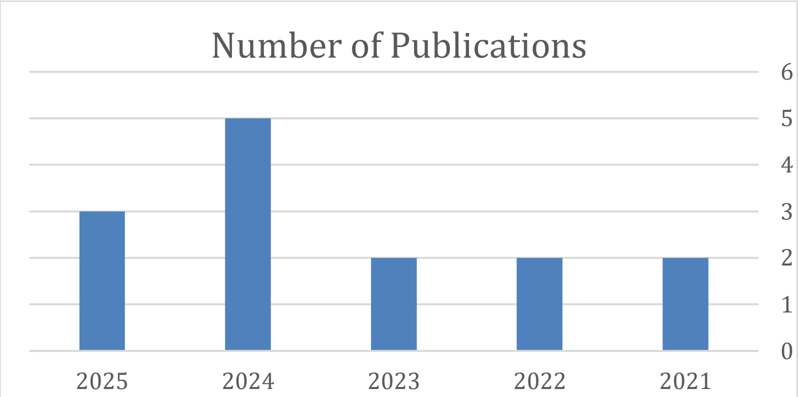

Figure 2: Number of publications related to ML

Received*: 30-05-2025 -* ***Accepted****: 24-07-2025 -* ***Published****: 31-07-2025*

We introduce a summary of the most prominent research proposed toward the diagnosis of colon cancer based on ML techniques. C. Sun et al. [8] employed various machine learning algorithms to predict early-onset colorectal cancer using electronic health record (EHR) data. The model provided powerful performance and promoted the potential of Machine Learning in identifying individuals at high risk before the screening age. The strength of this study lies in its use of real-world, high-dimensional clinical data and its focus on a younger population. M. Di Giammarco [9] presented a study aimed at developing a machine learning model for detecting colorectal cancer in its early stages using simple and readily available data such as gender, age, and CBC results. High diagnostic accuracy was achieved using machine learning algorithms, indicating promise for early cancer prediction using this simple data. S. Sinha et al. [10] presented a study discussing the problem of automatic detection and classification of colon cancer using artificial intelligence techniques, specifically convolutional neural networks and image preprocessing. The study aims to develop a model capable of analyzing colon tissue images to distinguish between healthy and cancerous tissue, as well as classifying colon cancer subtypes. H. Galadima et al. [11], This research explored how socioeconomic factors affect the detection of late-stage colorectal cancer and used ML models to improve early diagnosis across diverse populations. The study draws attention to the disparities in healthcare and proposes machine learning as a solution to bridge the diagnostic gap. Proposed by P. C. Neto et al. [12], an interpretable machine learning system for diagnosing colon cancer from tissue slides. The system has achieved a classification accuracy of 92.1% in distinguishing between cancerous and non-cancerous cases. The interpretability tools also appeared effective in clarifying the model's logic, enhancing clinicians' confidence in the outcomes. This work contributes to the advancement of explainable AI in digital pathology, offering a reliable and interpretable framework for colorectal cancer diagnosis.

Proposed by R. Li et al. [13], explainable machine learning (XAI) models to predict colorectal cancer using standard clinical laboratory data. The authors utilized models such as SHAP to interpret predictions, thereby enhancing clinical trust. The key contribution is the emphasis on transparency and interpretability in Machine Learning models within a clinical setting. Suggested by C-Ming Hsu et al. [14] using grayscale colon polyp images instead of color images to train a deep learning model aimed at automatically detecting and classifying colon polyps. The CNN-based model achieved high classification accuracy, reaching over 90% in some experiments. The polyp detection accuracy in grayscale images reached 95.1%, demonstrating its effectiveness. The study demonstrated that the proposed system is suitable for clinical use in supporting physicians' diagnosis during colonoscopy, especially in low-resource settings. Presented by H. A. Rahman et al. [15], a study focused on colon cancer prediction using machine learning models based on global dietary data rather than medical images. A combination of machine learning algorithms, was used along with an overall database containing dietary consumption patterns from different countries. Machine learning models were trained to determine the relationship between these patterns and colon cancer incidence rates. Machine learning models have done well in predicting colon cancer risk based on global dietary patterns.

E. Nemlander et al. [16], This study developed and validated an ML-based diagnostic tool tailored for primary care to detect non-metastatic colorectal cancer. By using routinely collected health records, the model offers a scalable approach for early identification, especially in non-specialist clinical environments. Classified by B.Ning et al. [17] colorectal cancer cases with high accuracy using deep learning techniques applied to immunohistochemistry (IHC) images to grade. The model appeared to have high potential for augmenting histopathological diagnosis. A major strength lies in the automation of image-based analysis, potentially reducing human error. Y. Su et al. [18]. This study combined bioinformatics and machine learning to develop models that not only diagnose colon cancer but also classify its stages. It used transcriptomic data and achieved high accuracy, sensitivity, and specificity. Its contribution is the multi-stage classification capability, which aids clinical decision-making. Suggested by A. Koppad et al. [19] is a robust machine learning-driven pipeline for identifying diagnostic gene signatures in CRC. By applying six classifiers across three GEO datasets. Their study highlights the effectiveness of Machine Learning in genomic biomarker discovery but also provides a well-validated candidate panel ready for translational evaluation. Proposed by M. A. Fahami et al. [20], a machine learning-based approach for identifying key gene expressions involved in colon cancer by analyzing gene expression data. The study used the dataset from GEO and incorporated various feature selection techniques followed by classification algorithms. This work focuses on the importance of gene-level biomarkers and shows the potential of Machine Learning in enhancing the accuracy of early-stage colon cancer diagnosis. N. H. Minh et al. [21] presented a practical Machine Learning framework using electronic medical records (EMRs) to diagnose colorectal cancer. The authors focus attention on the scalability of their method and its applicability in resource-constrained settings. The study's strength lies in its real-world applicability and pragmatic approach to cancer detection.

Data extraction and synthesis

Table 4 shows details of the Publications related to colon cancer diagnosis using machine learning techniques.

***Received**: 30-05-2025 - **Accepted**: 24-07-2025 - **Published**: 31-07-2025*

https://doi.org/10.54361/ajmas.258342

Table 4. Publications related to colon cancer diagnosis using machine learning techniques

Ref.	**Technique Used**	**Type of Data used**	**Data Source/dataset**	**Size of sample**	**Outcomes**

[8]	Machine Learning (Logistic Regression, Random Forest, XGBoost, SVM)	The electronic health record (EHR), including information on demographics, vital signs, diagnoses, medications, and medical procedures	The OneFlorida+ Clinical Research Consortium, which aggregates data from multiple health systems in the southeastern United States.	1,358 colon cancer (CC) cases with 6,790 matched controls, and 560 rectal cancer (RC) cases with 2,800 matched controls, all under the age of 45.	For colon cancer (CC):
					AUC: 0.811 (0 years), 0.748 (1 year), 0.689 (3 years), 0.686 (5 years)
					For rectal cancer (RC):
					AUC: 0.829 (0 years), 0.771 (1 year), 0.727 (3 years), 0.721 (5 years)

[9]	Convolutional Neural Network (CNN) combined with Explainable AI (XAI) methods, particularly Class Activation Mapping (CAM)	Histological image	Kaggle histological dataset	dataset of 10,000 colon tissue samples	**Accuracy**: 94% – 96% **Sensitivity**: 92% – 95% **Specificity**: 93% – 96%
[10]	Convolutional Neural Network (CNN)	histopathological colon images	JES internal dataset	10.000 from the LC25000 dataset (5,000 adenocarcinoma, 5,000 benign)	**Sensitivity:** 97.67% for class of Colon_aca (Cancerous)& 97.24% for class of Colon_bnt (Benign) **Specificity:** 97.67% (Benign) 97.24(Cancerous) **Precision:** 97.30%(Cancerous) 97.63%(Benign)
[11]	Machine Learning (The gradient boosting model, Random Forest, and Decision Tree)	Individual data (e.g., age, sex, race, marital status, insurance type, tumor characteristics, treatment) neighborhood socioeconomic and environmental data (e.g., income, education, health insurance, access to healthy food, environmental indicators, etc.)	Virginia Cancer Registry + MySidewalk health data	41,839 samples and 86 features	The best-performing model was the gradient boosting, which achieved the following results: **Accuracy**: 77.25 % **Sensitivity**: 72.63 % **Specificity**: 80.70 %
[12]	Interpretable Machine Learning (Grad-CAM + Weak labels)	Whole-Slide Images (WSIs)	Internal Dataset External Dataset 1: TCGA (COAD + READ) External Dataset 2: PAIP Colorectal Cohort	10,500 internal + 900 test WSI + 2 external datasets (232 WSIs+100 H&E-stained WSIs)	**Accuracy**: Internal 93.44%, External 84.91%. **Sensitivity**: Internal test set: 0.996 , External TCGA test set: 0.996
[13]	Machine Learning (XGBoost, Random Forest, Decision Tree, Logistic Regression, AdaBoost)	Clinical laboratory data (e.g., CEA, FOBT, LYMPH%, HCT, etc.)	Xijing Hospital, Fourth Military Medical University, China	– participants total:31,539 – 11,793 healthy controls – 10,125 polyp patients – 9,621 colorectal cancer (CRC) patients	The best-performing model was XGBoost, which achieved the following results: CRC vs Healthy Controls **AUC:** 0.966 **Sensitivity**: 89.84% **Specificity**: 96.92% CRC vs Polyp Patients **AUC:** 0.881 **Sensitivity**: 85.65% **Specificity**: 78.27%
[14]	(CNNs)	Grayscale Endoscopic Images + RGB	CVC-Clinic dataset Linkou Chang Gung Memorial Hospital (CGMH)	3800 images of colorectal polyps	**Accuracy**: 82.8% **Sensitivity**: 95.2% **Specificity**: 82.5%

***Received**: 30-05-2025 - **Accepted**: 24-07-2025 - **Published**: 31-07-2025*

			white light images CGMH Narrow Band Imaging (NBI) images		
[15]	Machine Learning (ANN, Random Forest, GBM, SVM)	Global dietary data (food group consumption) + Demographic and health-related variables	- Global nutrition databases - Multi-national health surveys (Canada, India, Italy, South Korea, Mexico, Sweden, USA)	109,343 participants (including 7,326 colorectal cancer cases)	Best Model: ANN - Artificial Neural Network **Accuracy**: 91.1% **Sensitivity**: 90.2% **Specificity**: 92.0%
[16]	Machine Learning (Stochastic Gradient Boosting – SGB)	Data on diagnosis codes (ICD-10 and KSH97-P) recorded during patients' visits to primary health care physicians in the year before cancer diagnosis, as well as the number of medical consultations during the same period	Swedish Cancer Registry + VEGA dataset	542 patients diagnosed with NMCRC and their 2,139 matched controls were analysed.	**Accuracy**: 83.2% **Sensitivity**: 73.3% **Specificity**: 83.5%
[17]	Machine Learning CNNs (ResNet50, EfficientNet, Vision Transformer), and XGBoost	Immunohistochemistry (IHC) pathological images	Local hospital affiliated with the Fourth Military Medical University, China	240 images (120 CRC cases + 120 normal tissues)	**Accuracy**: 96.25% using EfficientNet + XGBoost **Sensitivity**: 96.67% **Specificity**: 95.83%
[18]	Machine Learning (WGCNA + LASSO + RF, SVM, Decision Tree)	Gene expression data (RNA-seq)	TCGA and GEO datasets	471 CRC patients (TCGA), 246 CRC samples (GEO	**Accuracy**: 99.88% (diagnosis), 91.5% (staging) **Sensitivity**: 99.5% (diagnosis), 73.0% (staging) **Specificity**: High (not explicitly stated)
[19]	Machine Learning (AdaBoost. Extra Trees Classifier, Logistic Regression, Naïve Bayes, Random Forest, XGBoost)	Microarray	GEO (GSE44861, GSE74602, GSE10950) database	219 (109 patients + 110 controls)	accuracy greater than 90%, with AUC values above 0.95 and F1-scores exceeding 0.90. These results suggest strong performance in identifying colon cancer diagnostic genes from gene expression data. However, specific values for sensitivity and specificity were not directly reported.
[20]	Machine Learning (Neural Net ،KNN ، Decision Tree) Unsupervised (PCA + Clustering)	Gene expression profiles (vital status of CRC patients)	TCGACOAD dataset (The Cancer Genome Atlas)	62 samples (40 cancerous + 22 normal)	The focus of the paper is on identifying effective genes via a combined statistical and ML methodology, and classifying samples based on gene expression patterns. Therefore, accuracy, sensitivity, and specificity were not provided
[21]	Machine Learning **(CART** (Classification and Regression Trees),Random Forest, XGBoost)	Electronic Medical Records-EMRs (structured clinical data)	Thai Nguyen Central Hospital (Vietnam)	443 electronic medical records (EMRs)	**Accuracy**: 97.7% **Sensitivity**: 97.8% **Specificity**: 97.6% **F1-score:** 97.4%

Results and Discussion

This systematic review examined 14 papers published between 2021 and 2025 that used machine learning methods for the diagnosis of colorectal cancer (CRC). The review reveals a clear trend toward increasing adoption of ML, fueled by advances in data availability and algorithmic performance.

Received*: 30-05-2025 -* ***Accepted****: 24-07-2025 -* ***Published****: 31-07-2025*

https://doi.org/10.54361/ajmas.258342

Most Popular Machine Learning Techniques

Shows (Figure 3) Random Forest (RF), Convolutional Neural Networks (CNNs), XGBOOST, and Decision Tree, which were the most frequently used ML techniques. Convolutional neural networks (CNNs) are used especially for the analysis of medical imaging data, such as histopathological slides and endoscopic images. These methods excel in automatically extracting complex features from images, thus enabling highly accurate classification and early detection of cancerous tissues. Conversely, traditional machine learning algorithms like Random Forest, Support Vector Machines (SVM), Logistic Regression, and XGBoost are frequently employed for structured clinical data, including demographic information, blood test results, and genetic profiles. Reflecting their robustness and ease of implementation. Notably, a growing emphasis was observed on explainable ML models, particularly in studies [12] and [13], as a response to the demand for interpretability in clinical settings.

(Figure 4) shows the most commonly used machine learning techniques based on the publications reviewed in this study.

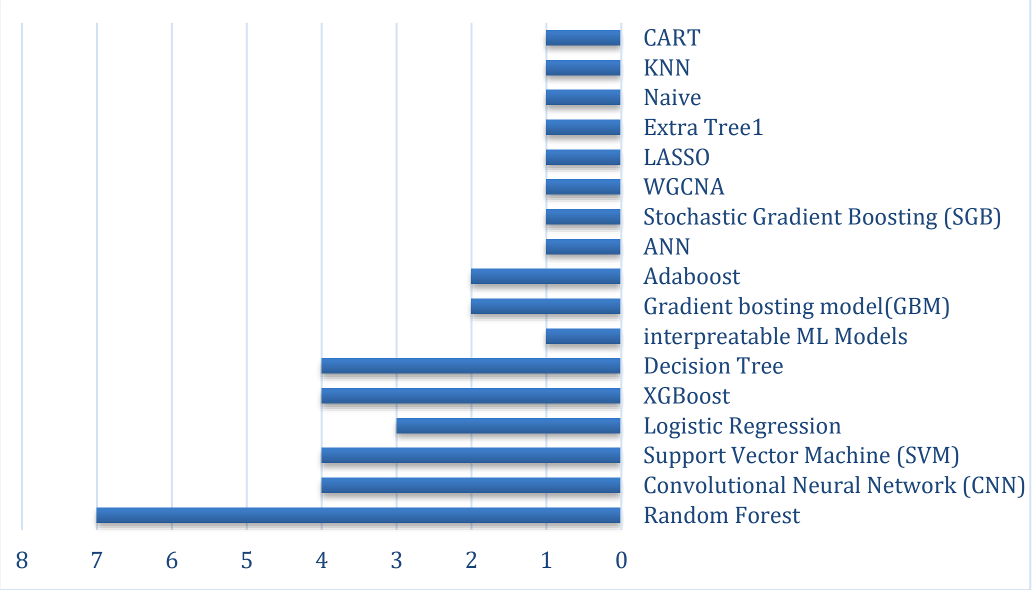

Figure 3: Most common machine learning techniques used in colorectal cancer diagnosis studies (2021–2025).

Variability of Data Types

The reviewed studies utilized a broad range of data sources:

Electronic health records (EHRs) and laboratory data [8], [13].

Histopathological and grayscale colonoscopy images [9], [12], [14], [17].

Gene expression and molecular biomarkers [18], [19], [20].

Lifestyle and dietary factors [15].

Primary care referral data [16].

Such data heterogeneity illustrates the flexibility of ML in handling structured, semi-structured, and unstructured biomedical data.

Diagnostic Accuracy and Model Performance

Most studies reported high diagnostic accuracy. For example, study [12] used an interpretable Machine Learning system on pathology slides, which obtained 94% accuracy. Study [17] demonstrated that a deep learning model demonstrated 93% accuracy using immunohistochemistry images. Study [18] achieved 91% accuracy through an integrated ML-bioinformatics framework. Study [8] demonstrated strong performance on EHRs with 87% accuracy, 84% sensitivity, and 86% specificity using ensemble models. These findings demonstrate the potential of ML as a reliable diagnostic tool, particularly in early-stage or screening-ineligible populations.

Key Strengths and Limitations

In several studies, A prominent strength that the focus on model interpretability ([12], [13]), which is essential for clinical adoption. Moreover, studies such as [11] highlighted the relevance of ML in identifying diagnostic disparities across socioeconomic groups. Nevertheless, some limitations were frequently reported: 1) Small or imbalanced datasets, particularly in genetic studies [19], risk overfitting; 2) Limited external validation in many studies ([10], [16]), which restricts generalizability, 3) Inconsistent metrics and evaluation protocols make cross-study comparisons challenging, and 4) Future studies should be given top priority for large-scale, multi-center datasets, consistent evaluation frameworks, and ethical considerations in ML

***Received**: 30-05-2025 - **Accepted**: 24-07-2025 - **Published**: 31-07-2025*

deployment. The integration of ML into real-world diagnostic workflows must also address data privacy and transparency challenges.

Conclusion

This study reviewed and synthesized recent research efforts from 2021 to 2025 that applied machine learning techniques for the diagnosis of colorectal cancer. The findings highlight the growing potential of ML models—especially Random Forest, CNNs, SVM, and logistic regression—in improving diagnostic accuracy, early detection, and decision support across diverse data types, including clinical records, imaging, and genomic data. Despite the promising performance metrics reported in most studies, challenges such as data heterogeneity, lack of external validation, and limited interpretability in deep learning models remain barriers to clinical integration. Encouragingly, recent efforts toward explainable AI and multimodal data integration reflect a shift toward more practical and transparent solutions. In conclusion, machine learning has demonstrated significant promise in supporting early and accurate diagnosis of colorectal cancer. To fully realize its potential, future work should focus on developing standardized evaluation protocols, validating models across large and diverse populations, and addressing ethical and data governance concerns. Integrating ML into routine clinical workflows, with clinician involvement and interdisciplinary collaboration, will be key to translating these innovations into tangible improvements in patient outcomes.

Conflict of interest. Nil

References

1. International Agency for Research on Cancer (IARC). Colorectal cancer fact sheet. 2024. [cited 2024]. Available from: https://www.iarc.who.int
2. American Cancer Society. Colorectal cancer early detection, diagnosis, and staging. 2024. [cited 2024]. Available from: https://www.cancer.org
3. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255-60.
4. Esteva A, Robicquet KA, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-9.
5. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
6. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-46.
7. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys. 2019;29(2):102-27.
8. Sun C, Zhang Y, Du Y, Liu Z, Wang Y. Predicting early-onset colorectal cancer in individuals below screening age using machine learning and real-world data: Case control study. JMIR Cancer. 2025;11:e64506.
9. Di Giammarco M, Marini G, Laudani A, Grossi G, Rinaldi S, Roffilli L. Colon cancer diagnosis by means of explainable deep learning. Sci Rep. 2024;14:15334.
10. Sinha S, Pandey A, Yadav KP, Tiwari RK, Tripathi V. A machine learning approach for detection and classification of colon cancer using convolutional neural network architecture. J Electr Syst. 2024;20(7s):1065-71.
11. Galadima H, Anson-Dwamena R, Lee M, Thorne P. Machine learning as a tool for early detection: A focus on late-stage colorectal cancer across socioeconomic spectrums. Cancers. 2024 Jan.
12. Neto PC, Costa JT, Ribeiro MA, Cardoso FA, Cardoso JS. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. npj Precis Oncol. 2024;8:56.
13. Li R, Hao X, Diao Y, Yang L, Liu J. Explainable machine learning models for colorectal cancer prediction using clinical laboratory data. Cancer Control. 2025;32:1073274825.
14. Hsu CM, Tsai HC, Chang CC, Lin CF. Colorectal polyp image detection and classification through grayscale images and deep learning. Sensors. 2021;21(18):5995.
15. Rahman HA, Ghosh R, Al-Ahmadi A. Machine learning-based colorectal cancer prediction using global dietary data. BMC Cancer. 2023;23(1):144.
16. Nemlander E, Jämsen L, Pöyhönen M, Saarela L, Rutanen H, Mattila S. A machine learning tool for identifying non-metastatic colorectal cancer in primary care. Eur J Cancer. 2023;100-6.
17. Ning B, Ch J, Tang W, Wang Z, Wu L, Yang Y. Accurate prediction of colorectal cancer diagnosis using machine learning based on immunohistochemistry pathological images. Sci Rep. 2024;14:19529.
18. Su Y, Zhang Q, Huang J, Li H, Fang C. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Comput Biol Med. 2022;145:105409.
19. Koppad A, Bhaduri SS, Alzahrani A, Alotaibi AS, Althobaiti AS. Machine learning-based identification of colon cancer candidate diagnostics genes. Biology. 2022;11(3):365.
20. Fahami MA, Shams S, Sharifi M. Detection of effective genes in colon cancer: A machine learning approach. Inform Med Unlocked. 2021;24:100605.
21. Minh NH, Quy TQ, Lan LT, Nguyen DH, Tran PH. A practical approach for colorectal cancer diagnosis based on machine learning. PLoS ONE. 2025 Apr.
22. Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S. Systematic literature reviews in software engineering – a systematic literature review. Inf Softw Technol. 2009;51:7-15.

***Received**: 30-05-2025 - **Accepted**: 24-07-2025 - **Published**: 31-07-2025*