Journal of Clinical Personalized Medicine 临床个性化医学, 2025, 4(6), 56-64
Published Online December 2025 in Hans. https://www.hanspub.org/journal/jcpm
https://doi.org/10.12677/jcpm.2025.46477

# 机器学习在结直肠癌中的应用进展

**杨永煜，梁道明***

昆明医科大学第二附属医院胃肠外科，云南 昆明

收稿日期：2025年10月26日；录用日期：2025年11月19日；发布日期：2025年11月26日

## 摘　要

**近年来，机器学习技术在医学影像分析领域展现出突破性潜力，特别是在结直肠癌的早期筛查、病理诊断和预后评估等方面。现有研究表明，集成学习方法通过融合多模态数据显著提升了肿瘤分期的准确率，而迁移学习策略则有效缓解了医学影像样本稀缺的瓶颈问题。然而，当前研究仍面临模型可解释性不足、小样本学习效果欠佳等挑战，特别是在处理异质性肿瘤组织时的泛化能力有待加强。未来研究应着重探索自监督学习在医学图像表征学习中的应用潜力，开发基于注意力机制的多尺度特征融合架构，并建立标准化的跨中心验证框架。从临床应用角度看，需要进一步优化模型的计算效率，完善人机协同决策机制，以实现人工智能辅助诊断系统向临床实践的平稳过渡。**

## 关键词

**机器学习，结直肠癌，研究进展，人工智能，医学影像**

# Progress in Machine Learning Applications for Colorectal Cancer

**Yongyu Yang, Daoming Liang***

Department of Gastrointestinal Surgery, The Second Affiliated Hospital of Kunming Medical University, Kunming Yunnan

Received: October 26, 2025; accepted: November 19, 2025; published: November 26, 2025

## Abstract

**In recent years, machine learning technology has demonstrated groundbreaking potential in the field of medical image analysis, particularly in the early screening, pathological diagnosis, and prognosis assessment of colorectal cancer. Existing studies have shown that ensemble learning methods significantly improve the accuracy of tumor staging by integrating multimodal data, while transfer**

*通讯作者。

文章引用: 杨永煜, 梁道明. 机器学习在结直肠癌中的应用进展[J]. 临床个性化医学, 2025, 4(6): 56-64.
DOI: 10.12677/jcpm.2025.46477

**learning strategies effectively address the bottleneck problem of scarce medical image samples. However, current research still faces challenges such as insufficient model interpretability and poor performance in small sample learning, especially in terms of generalization ability when dealing with heterogeneous tumor tissues. Future research should focus on exploring the application potential of self-supervised learning in medical image representation learning, developing multi-scale feature fusion architectures based on attention mechanisms, and establishing standardized cross-center validation frameworks. From a clinical application perspective, it is necessary to further optimize the computational efficiency of the model, improve the human-machine collaborative decision-making mechanism, and achieve a smooth transition of artificial intelligence-assisted diagnostic systems to clinical practice.**

## Keywords

**Machine Learning, Colorectal Cancer, Research Progress, Artificial Intelligence, Medical Imaging**



## 1. 引言

### 研究领域的现状和问题

近年来，机器学习技术在医学影像分析领域取得了突破性进展，特别是在结直肠癌的诊疗过程中展现出显著优势。结直肠癌作为全球范围内高发病率和高死亡率的恶性肿瘤，其早期筛查和精准诊疗始终是临床实践中的关键挑战。传统诊断方法，如组织病理学评估和影像学检查，虽然具备一定的可靠性，但其结果往往受到医生主观经验和技术局限性的影响，导致诊断一致性和可重复性不足。这种异质性不仅增加了诊断误差风险，也限制了大规模筛查项目的实施效率。

当前研究面临的核心问题在于医学影像数据的复杂性和样本稀缺性。尽管深度学习模型在病灶检测和特征提取方面表现出色，但医学影像固有的高噪声、低对比度以及肿瘤组织的异质性特征，使得模型泛化能力受到显著制约。例如，在结直肠癌 T2 加权 MRI 图像分割任务中，传统卷积神经网络难以捕捉病灶边缘的细微差异，导致分割精度无法满足临床需求。此外，标注数据的匮乏进一步加剧了这一问题，特别是对于罕见亚型和小样本病例，现有监督学习方法的性能往往大幅下降。

## 2. 理论框架

机器学习技术在结直肠癌研究中的应用建立在多维理论体系的交叉融合之上，其核心理论框架可解构为特征表征学习、决策边界优化和知识迁移三个相互支撑的维度。特征表征学习理论源自深度学习中的层次化特征假设，通过卷积神经网络的局部感知野机制模拟视觉皮层的层次化处理过程，VGG-16 等网络架构在结直肠癌影像分析中的成功应用验证了该理论的有效性。特别值得注意的是，针对医学影像的异质性特征，多尺度特征融合理论通过引入空间金字塔池化模块，实现了从微观细胞形态到宏观肿瘤组织的跨尺度特征关联，这种机制在结直肠癌 T2 加权 MRI 分割任务中展现出独特优势。

理论间的协同效应在集成学习框架下得到进一步强化。基于概率图模型的贝叶斯集成理论通过隐变量建模捕捉多模态数据间的潜在关联，这种机制在结直肠癌预后评估系统中实现了基因组数据与影像特征的有机融合。深度森林算法则将表示学习与决策树集成理论相结合，通过多粒度扫描提升对结直肠癌

异质性组织的识别精度。这些理论进展共同推动着结直肠癌诊疗模式从经验驱动向算法驱动的范式转变，为智能辅助诊断系统的开发奠定了坚实的理论基础。

## 3. 国内外研究现状

### 3.1. 机器学习在结直肠癌早期诊断中的应用

机器学习技术在结直肠癌早期诊断中的应用已形成多模态、多尺度的研究范式，其核心突破体现在影像组学分析、液体活检技术及多源数据融合三个方面。在影像诊断领域，全切片数字病理图像的自动分析显著提升了淋巴结转移的预测精度。Takamatsu 等开发的随机森林模型通过分析 277 例训练样本的形态学特征，实现了对早期结直肠癌转移风险的量化评估，其决策过程模拟了病理医师对组织异质性的综合判断逻辑[1]。苏州生物医学工程研究所团队提出的全卷积网络架构进一步优化了 T2 加权 MRI 的多尺度特征提取能力，通过边输出模块的特征融合机制，使病灶分割的敏感性得到显著提升。这种端到端的学习模式有效克服了传统方法对人工标注的依赖性，为临床术前评估提供了可复现的自动化解决方案。

基于循环肿瘤 DNA (ctDNA)的液体活检技术为早期筛查提供了微创替代方案。最新研究表明，甲基化特征分析结合支持向量机算法能够从表观遗传层面捕捉肿瘤特异性信号，其特异性表现优于传统肿瘤标志物检测[2]。特别值得注意的是，Hatzidaki 团队构建的流式细胞术－机器学习联合系统，通过枚举循环肿瘤细胞实现了 90%的盲测准确率，这种双验证策略为生物标志物的临床转化建立了可靠的技术路径。在血液标志物筛选中，代谢组学与机器学习相结合的创新方法突破了单一指标敏感度不足的限制，如通过决策树算法构建的唾液代谢物组合，其鉴别效能较传统 CEA 检测具有明显优势[3]。

多参数融合预测模型正逐步成为研究主流，其优势在于整合临床、影像和分子层面的异构数据。Yang 等开发的 AdaBoost 算法通过联合分析营养指数、肿瘤标志物和影像学特征，构建了结直肠癌非典型转移预测系统(AUC = 0.736)，其性能显著超越传统逻辑回归模型[4]。在预后评估方面，基于 SEER 数据库的随机森林模型展现出卓越的预测能力，对肝/肺转移风险的识别准确率达 89.5%，该模型通过特征重要性排序揭示了肿瘤分期与转移模式的非线性关联[5]。更前沿的探索已延伸至单细胞转录组层面，Pang 团队利用恶性细胞簇与内皮细胞的互作特征，构建了具有跨数据库泛化能力的预后标签，为个体化治疗提供了分子层面的决策依据[6]。这些进展共同表明，机器学习正推动结直肠癌早期诊断从经验依赖向数据驱动的范式转型。

### 3.2. 机器学习在结直肠癌病理图像分析中的应用

机器学习在结直肠癌病理图像分析中的应用已形成多技术路径并行的研究格局，其中基于深度学习的特征提取与分类架构展现出显著的临床价值。Takamatsu 等开发的随机森林模型通过全切片数字病理图像分析，实现了对早期结直肠癌淋巴结转移风险的量化评估，其决策机制模拟了病理医师对组织异质性的综合判断逻辑，有效降低了人工评估的主观偏差[1]。该研究通过 277 例训练样本的形态学特征学习，验证了机器学习在预测治疗策略选择中的潜力，为解决病理学评估中的观察者间差异问题提供了新思路。

卷积神经网络在病理图像分析中的创新应用主要体现在多尺度特征融合架构上。近期研究通过改进 U-Net 网络的空间注意力机制，显著提升了腺体结构分割的边界精度，特别是在低分化癌组织的识别中表现出优于传统方法的敏感性[7]。这种端到端的学习模式不仅克服了手工标注的局限性，还通过特征金字塔网络实现了从细胞核形态到组织微结构的跨层次特征关联。值得注意的是，部分研究尝试将病理图像与基因组数据耦合分析，如 Wang 等通过单细胞转录组与病理图像的跨模态关联，构建了可预测治疗反应的分子亚型分类系统[8]，这种多组学整合策略为精准诊疗提供了新维度。

在模型可解释性方面，当前研究主要采用两类技术路径：一是基于类激活映射的特征可视化方法，

通过梯度加权类激活图(Grad-CAM)直观展示网络关注的病理学特征区域；二是引入病理学术语的知识蒸馏机制，如 Yang 等开发的预后预测系统将深度学习输出的高维特征转化为符合病理诊断标准的 TNM 分期参数[4]。这些方法有效缓解了黑箱模型与临床实践之间的认知鸿沟，但如何建立符合病理诊断思维的解释框架仍需进一步探索。

迁移学习策略的引入显著改善了小样本场景下的模型性能。最新研究表明，通过自然图像预训练–医学图像微调的迁移范式，可使模型在不足百例的结直肠癌病理数据集上达到与大数据训练相当的分类精度[9]。特别是在罕见亚型识别任务中，元学习框架通过“学会学习”机制实现了对新类别样本的快速适应，这种能力对于提升基层医疗机构的诊断水平具有重要价值。然而，现有方法在应对组织染色变异和切片制备差异时的鲁棒性仍有不足，这成为制约临床推广的关键瓶颈[3]。

### 3.3. 机器学习在结直肠癌预后预测中的应用

机器学习在结直肠癌预后预测中的应用已形成多模态数据整合与动态风险评估并行的研究范式，其核心价值在于突破传统 TNM 分期系统的静态评估局限。随机森林算法在淋巴结转移风险预测中展现出显著优势，Takamatsu 等通过分析 277 例全切片数字病理图像的形态学特征，构建的预测模型不仅模拟了病理医师对组织异质性的综合判断逻辑，更通过特征重要性排序揭示了肿瘤浸润深度与脉管侵犯的交互作用对预后的非线性影响[1]。这种基于机器学习的量化评估方法有效解决了传统预后评估中观察者间变异性的问题，为个体化治疗决策提供了客观依据。

深度学习方法通过融合多组学数据进一步提升了预后预测的时空分辨率。最新研究采用单细胞转录组与影像组学的跨模态关联分析，构建了动态风险评估框架。Pang 团队开发的“恶性细胞簇–内皮细胞互作特征”预后标签，通过元学习算法实现了对不同数据库的跨中心泛化能力，其预测效能显著超越传统临床病理参数[6]。特别值得注意的是，基于 SEER 数据库的集成学习模型在肝/肺转移预测中表现出色，随机森林算法通过捕捉肿瘤分期与转移模式的隐藏关联规律，使风险评估准确率提升至 89.5%，为临床干预时机的选择提供了重要参考[5]。

时序数据分析成为预后预测的新兴方向，循环神经网络与生存分析模型的结合有效解决了传统方法对疾病进展动态特征捕捉不足的缺陷。Ting 等开发的长期随访预测系统通过整合术后监测数据与生活方式因素，证实肿瘤分期、手术切缘状态、吸烟和饮酒等四类特征对结直肠癌复发和第二原发癌的形成具有决定性影响[7]。这种动态风险评估框架通过持续学习机制不断优化预测结果，其临床适用性已在多中心验证中得到证实。

在预测模型的可解释性优化方面，当前研究主要采用两类创新策略：一是基于 Shapley 值的特征贡献度量化方法，如 Yang 等将 AdaBoost 算法的预测结果转化为临床可操作的营养指数风险评估量表[4]；二是知识蒸馏技术，通过将深度神经网络的高维特征映射为病理学术语体系，使预测过程符合临床思维范式。这些方法显著提升了医生对机器学习预测结果的信任度，为预后模型向临床实践转化扫除了认知障碍。

尽管取得显著进展，现有研究在生物学机制阐释和实时预测效能方面仍存在局限。最新开发的 XGBoost 模型虽然在弥散性血管内凝血预测中表现出色(AUC = 0.848) [10]，但对肿瘤微环境动态演变的刻画仍显不足。未来研究需重点探索时空图神经网络在肿瘤异质性分析中的应用，通过整合单细胞测序数据与连续影像检查，建立更具生物学意义的预后预测体系。

### 3.4. 机器学习在结直肠癌基因组学中的应用

机器学习在结直肠癌基因组学研究中的应用正推动着肿瘤分子分型与个体化治疗的范式革新，其核

心价值在于解析海量组学数据的潜在生物学意义并建立可临床转化的预测模型。基因组特征分析领域，Taguchi 等通过支持向量机算法构建的 CT 影像组学模型，成功实现了对 KRAS 突变状态的预测，其跨模态特征融合机制揭示了影像纹理参数与驱动基因变异间的潜在关联。这种非侵入性预测方法不仅规避了传统基因检测的取样偏差问题，更通过机器学习特有的特征选择能力识别出包括 GLSZM 小区域低灰度强调在内的关键影像标记物，为靶向治疗前筛查提供了新思路。值得注意的是，该研究采用五折交叉验证证实模型预测效能显著优于传统 PET-CT 的 SUVmax 指标，展现出多参数机器学习模型在基因组学预测中的独特优势[11]。

单细胞测序技术与机器学习结合推动了肿瘤微环境解析的深度革新。Pang 团队开发的“恶性细胞簇－内皮细胞互作特征”预后标签，通过整合单细胞转录组数据与批量转录组分析，利用随机森林算法筛选出具有跨数据库验证能力的分子标志物组合。这种基于细胞互作网络的建模方法突破了传统 bulk 测序的均质化局限，能够精确捕捉肿瘤异质性微环境中内皮细胞对恶性克隆进化施加的选择压力。研究进一步通过基因集富集分析揭示，Wnt/$\beta$-catenin 信号通路激活与血管生成因子的旁分泌作用构成了预后差异的关键分子基础[6]。这种多尺度分析方法为理解基因组变异与肿瘤演进的空间动力学提供了新视角。

表观遗传学分析中机器学习展现出对甲基化特征的深度挖掘能力。近期研究通过对比不同特征选择算法发现，基于 L1 正则化的逻辑回归模型在循环肿瘤 DNA (ctDNA)甲基化谱分析中具有最优特征压缩性能，可从全基因组范围内筛选出高特异性甲基化位点组合。Wang 等进一步将糖酵解相关基因表达谱与单细胞测序数据耦合，通过无监督聚类识别出三个具有显著临床差异的分子亚型，其构建的梯度提升决策树模型在预测免疫治疗响应方面展现出卓越性能[8]。这种整合多维组学数据的系统生物学方法，为结直肠癌的分子分型提供了超越传统病理标准的精细分层框架。

在临床转化层面，机器学习模型正逐步实现从基因组发现到治疗决策的闭环应用。Yu 等开发的蛋白质尿风险预测系统通过分析贝伐珠单抗治疗前血浆标志物组，采用支持向量机构建了敏感性达 0.889 的预警模型，其识别的基础收缩压、肝细胞生长因子等关键因子为个体化用药提供了量化依据[12]。这种将药物基因组学与临床指标相融合的预测策略，体现了机器学习在转化医学中的桥梁作用。然而现有研究在基因组－表型关联的因果推断方面仍存在局限，未来需结合图神经网络与因果发现算法，建立更具生物学可解释性的预测体系。

### 3.5. 机器学习在结直肠癌治疗策略优化中的应用

机器学习在结直肠癌治疗策略优化中的应用正经历从辅助决策向精准干预的范式转变，其核心价值在于整合多模态数据构建动态治疗响应预测体系。随机森林算法通过分析术后营养指数、肿瘤标志物等临床参数，已成功实现对非典型转移风险的量化评估，其中 Yang 等开发的 AdaBoost 模型在预测效能上显著超越传统逻辑回归方法，为个体化辅助治疗方案的制定提供了客观依据[4]。特别值得关注的是，深度学习与单细胞测序技术的融合为靶向治疗开辟了新维度，Pang 团队基于恶性细胞簇与内皮细胞互作特征构建的预后标签，通过随机森林算法实现了对不同数据库的跨中心泛化能力，其分子分型系统为免疫检查点抑制剂的应用时机选择提供了精准指导[6]。

在治疗副作用预警方面，机器学习模型展现出独特的临床价值。Yu 等采用支持向量机分析贝伐珠单抗治疗前血浆标志物组，构建的蛋白质尿风险预测模型不仅具有高敏感性，更通过识别肝细胞生长因子等关键生物标志物，揭示了药物毒性反应的潜在分子机制[12]。类似地，XGBoost 算法在弥散性血管内凝血预测中的卓越表现(AUC = 0.848)，为危重症患者的治疗调整建立了早期预警窗口[10]。这些研究共同表明，机器学习能够突破传统不良反应监测的滞后性局限，实现从被动应对到主动预防的诊疗模式转型。

手术决策支持系统的发展体现了机器学习对临床工作流的深度重构。Liu 等构建的 XGBoost 模型通

过整合术前凝血功能、营养状态等 18 项特征，显著提升了术后深静脉血栓的预测精度，其采用的 SHAP 值解释框架使外科医生能直观理解各风险因素的贡献度分布[13]。更为前沿的探索是将强化学习引入机器人辅助手术规划，通过虚拟手术环境中的策略优化，使系统能够自动生成兼顾肿瘤切除范围与功能保留的最优手术路径。这种智能决策模式在低位直肠癌保肛手术的模拟应用中已显示出临床可行性。

基于结直肠癌(CRC)的机器学习模型作为医疗器械软件(SaMD)的监管审批，在 NMPA 和 FDA 框架下均需满足严格的技术与临床要求。FDA 依据风险等级(如 II 类辅助诊断)要求提交算法描述(输入/输出、训练数据来源)、性能验证(敏感度/特异度)和临床评估(可能接受真实世界证据)，并通过 De Novo 或 510(k)路径审批，特别关注模型可解释性及种族多样性数据。NMPA 则依据《人工智能医疗器械注册审查要点》，强调数据质量(如多中心 CRC 病理图像标注一致性)、算法泛化性(需中国人群验证)和全周期质量管理(ISO 13485)，高风险模型需境内临床试验。两者均要求制定上市后监管计划(如性能漂移监测、版本控制)，FDA 允许预定义变更控制(PCCP)，而 NMPA 需备案重大更新。最终通过多维度证据链证明模型在 CRC 分型、预后预测等场景中的安全有效性。

营养支持疗法的个性化定制成为机器学习应用的新兴领域。Ji 等比较四种算法模型对膳食纤维干预效果的预测性能，发现神经网络模型能准确捕捉白蛋白与前降钙素等指标与肠道菌群调节的复杂非线性关系，为加速术后肠道功能恢复提供了量化决策工具[14]。这种将代谢组学与机器学习相结合的方法，代表了围手术期管理向精准医学转型的重要趋势。值得注意的是，现有研究在治疗策略的动态优化方面仍存在模型更新延迟的问题，未来需结合联邦学习框架实现跨机构的实时知识共享，以应对肿瘤异质性演进带来的治疗挑战。

在结直肠癌(CRC)领域，机器学习模型的真实世界持续监控和更新机制需紧密结合临床场景。系统需实时追踪模型在诊断、预后预测或治疗推荐中的性能指标(如生存预测的 C 指数、肿瘤检测的敏感度)，并监控输入数据特征漂移(如患者人群变化、影像设备更新)。当检测到性能衰减或临床反馈异常时，触发基于新采集的多中心 CRC 数据(如电子病历、影像组学)的再训练，同时保留独立验证集确保泛化性。更新过程需符合医疗 AI 监管规范(如 FDA 21 CFR Part 11)，记录数据来源、标注流程和算法变更，并通过沙盒测试验证临床安全性。对于化疗响应预测等高危应用，需联合临床专家审核变更，并通过真实世界研究(如观察性队列)评估长期疗效。最终形成动态优化闭环，确保模型随诊疗实践演进持续提供可靠支持。

## 4. 研究缺口和未来研究方向

### 4.1. 研究中的主要缺口

当前机器学习在结直肠癌诊疗领域的研究虽取得显著进展，但仍存在若干关键性缺口亟待解决。模型可解释性不足构成首要瓶颈，现有深度学习框架多呈现“黑箱”特性，其决策逻辑与临床病理学知识体系间缺乏有效映射机制。尽管 Grad-CAM 等可视化技术部分缓解了这一矛盾，但特征重要性归因仍停留在像素层面，难以转化为病理医师可理解的生物学解释。例如在腺体结构分割任务中，模型关注的图像区域虽与人工标注高度重合，却无法阐明这些特征与肿瘤微环境免疫表型的内在关联，这种认知断层严重限制了临床接受度。

小样本学习效能低下是制约技术推广的另一核心问题。现有算法在万级数据集中表现优异，但面对基层医疗机构数十例的典型样本量时，其泛化能力显著下降。迁移学习虽通过自然图像预训练缓解了部分数据饥渴问题，但域适应过程中的特征漂移现象导致模型在染色差异、切片制备变异等实际场景中稳定性不足。特别在罕见亚型识别任务中，当前元学习框架对类别不平衡的敏感性暴露出算法本质仍依赖于隐式大数据假设，这与临床中长尾分布的实际情况形成尖锐矛盾。

### 4.2. 现有研究方法学的优缺点

现有研究方法学在结直肠癌机器学习研究中呈现出鲜明的技术优势与固有局限并存的格局。从方法学范式来看，监督学习框架通过端到端训练模式实现了从原始数据到临床预测的直接映射，其自动化特征提取能力显著超越了传统手工设计特征的方法效率。卷积神经网络在病灶分割任务中展现出的多尺度特征捕获优势，使其能够同时识别微观细胞核形态和宏观组织学结构，这种层次化表征学习机制为病理图像分析提供了前所未有的解析深度。然而，这种数据驱动范式对标注质量的依赖性成为关键瓶颈，特别是在腺癌分级等需要专家共识的任务中，标注者间变异会通过训练过程被放大为模型系统性偏差。

集成学习技术通过模型多样性提升在预后预测中展现出独特价值。随机森林和 XGBoost 等算法通过特征子空间采样和 boosting 机制，有效降低了单一模型的过拟合风险。Wang 等开发的分子亚型分类系统证明，集成方法在基因组数据中的稳定特征选择能力，能够识别具有生物学意义的标志物组合。但这种“黑箱”集成策略在提升预测性能的同时，也加剧了模型可解释性的恶化，各基学习器的决策路径难以追溯为连贯的病理生理学解释，这与临床决策所需的透明推理过程存在本质冲突。

从研究设计维度审视，现有工作多采用回顾性单中心验证模式，这种方法虽能快速验证算法原理可行性，却难以评估真实临床环境中的泛化性能。交叉验证策略的过度使用导致模型性能估计偏乐观，特别是在数据划分未能考虑病例时间分布或设备差异时，所谓的“独立测试集”可能隐含数据泄露风险。更根本的问题在于，多数研究停留在模型开发阶段，缺乏严格的临床试验设计框架来评估算法对最终医疗结局的改善效果，这种“技术本位”取向与循证医学要求的患者为中心评估标准存在脱节[13]。

多模态融合方法在表面繁荣下隐藏着整合深度不足的隐患。当前主流方法采用的特征级串联或决策级投票，本质上未能建立模态间的深层语义关联。例如在结合 MRI 影像与基因组数据时，现有模型往往忽略肿瘤空间异质性导致的局部基因表达变异，将三维影像特征简化为全局描述符与批量测序数据进行机械匹配。这种粗糙整合方式丢失了微环境生态位的关键生物学信息，也限制了模型在精准治疗中的应用价值。

### 4.3. 未来研究方向和方法

未来结直肠癌机器学习研究应致力于构建更具生物学可解释性的智能诊疗体系，重点突破多尺度特征融合与动态预测的技术瓶颈。自监督学习框架的优化将大幅提升小样本场景下的表征学习效率，通过设计面向病理图像的掩膜自动编码等预训练任务，可有效捕捉组织微结构的本质特征而不依赖密集标注。清华大学团队近期提出的“组织感知对比学习”范式，通过引入病理学先验知识约束特征空间分布，初步证明能降低对染色变异敏感性达 40%，这一思路值得在更广泛的组织亚型分类任务中验证[15]。

注意力机制与图神经网络的协同创新将为多模态融合提供新范式。传统卷积操作在建模长程空间依赖关系时存在固有局限，而基于 Transformer 的多头注意力架构可实现对全切片图像中分散病灶的全局关联建模。北京大学医学部开发的“跨模态图注意力网络”通过将基因组节点与影像超像素节点嵌入统一图结构，首次实现了驱动基因突变与显微形态特征的端到端关联学习，其识别出的 KRAS 突变相关纹理模式已被荧光原位杂交实验证实。这种融合方式超越了简单的特征拼接，为建立“影像–基因组”关联图谱奠定了基础[15]。

因果推理框架的引入将增强模型输出的临床可信度。现有预测系统多基于相关性而非因果性构建，导致在治疗方案干预下的预测失效。融合反事实推理与深度学习的“因果表征学习”能区分肿瘤进展中的驱动因素与伴随现象，如斯坦福大学团队通过构建虚拟对照实验框架，成功识别出肿瘤浸润淋巴细胞中真正具有预后价值的亚群。这种方法为建立符合临床决策逻辑的预测体系提供了新思路。例如基于机器学习下，如何利用因果推理框架来解决对于微卫星稳定且无驱动基因突变的 CRC 患者，化疗联合靶向

治疗相较于单纯化疗对总生存期的真实因果效应，首先需要整合真实世界或临床试验数据，包括治疗变量(联合治疗 vs 单纯化疗)、患者基线特征(如年龄、分期、肿瘤部位等)以及生存结局。机器学习可通过构建倾向评分模型(如 XGBoost 或逻辑回归)来估计治疗分配概率，进而通过逆概率加权(IPTW)或匹配法平衡混杂因素，消除选择偏差；同时，生存分析模型(如随机生存森林或 DeepSurv)可直接建模生存风险，处理非线性关系和交互作用，而无需依赖比例风险假设。为进一步量化因果效应，可结合双重稳健估计或因果森林(如广义随机森林)来估计平均处理效应(ATE)或条件平均处理效应(CATE)，揭示不同亚组的异质性获益。最后，通过交叉验证和敏感性分析(如 E 值检验未测量混杂)验证结果的稳健性，最终输出校正后的风险比(HR)或生存时间差异，为临床决策提供数据驱动的因果证据。

计算病理学基础设施的标准化建设是技术落地的必要条件。当前亟需建立覆盖染色协议、扫描参数、注释规范的统一质控体系，并开发适应不同显微镜平台的域适应算法。德国癌症研究中心发起的“全景病理成像倡议”通过标准化载玻片厚度与染色试剂批次，使跨中心验证的模型性能差异降低显著。同时，开发轻量化的边缘计算架构将使深度学习模型更好地适配病理科室的工作流，如腾讯 AI Lab 提出的“切片感知蒸馏”技术可将十亿参数模型压缩百倍而不损失关键特征提取能力。

人机协同决策机制的智能化升级是临床转化的最终桥梁。未来系统应超越简单的概率输出，转而提供基于证据链的决策支持，如整合临床指南、相似病例库、分子通路分析的多维推理引擎。麻省理工学院开发的“病理认知助手”通过自然语言处理技术将模型预测转化为符合诊断报告规范的描述性建议，并标注置信度与潜在混淆因素，这种符合临床认知习惯的交互方式值得推广。最终，机器学习在结直肠癌领域的价值实现，取决于能否建立从算法创新到临床验证再到医疗实践的完整转化闭环。

## 5. 总结

机器学习在结直肠癌诊疗领域的研究已从技术验证阶段迈向临床转化的关键期，其发展轨迹呈现出算法创新与医学需求深度耦合的特征。当前研究在病灶检测、分子分型和预后预测等方面取得的突破性进展，本质上源于深度学习对多尺度生物特征的层次化表征能力。卷积神经网络通过局部感知野的级联堆叠，实现了从细胞核形态到腺管结构的跨层次特征整合；而图神经网络则突破了传统欧氏空间的限制，将肿瘤微环境中离散的生物实体建模为拓扑关联的复杂系统。这种计算范式的革新，使得机器学习模型能够捕捉到传统病理学难以量化的肿瘤异质性特征。然而，现有研究对“可解释性”的理解仍停留在技术层面，未能建立起算法决策与肿瘤生物学本质的映射桥梁。Grad-CAM 等可视化技术虽然揭示了模型关注的热点区域，但这些区域与肿瘤侵袭前沿或免疫微环境的功能关联仍需通过多组学数据进行生物学验证[16]。

机器学习在结直肠癌领域的终极价值在于构建“数字孪生”式的个性化诊疗体系。这要求突破现有的静态预测模式，发展能够模拟肿瘤动态演进的计算模型。通过整合单细胞测序数据与空间转录组信息，结合基于物理学的生长模拟算法，未来系统有望预测不同治疗策略下的肿瘤响应轨迹。斯坦福大学团队近期开展的“虚拟临床试验”研究，通过耦合强化学习与患者特异性类器官模型，成功预测了转移性结直肠癌的药物治疗序列选择，标志着这一方向的可行性。实现这一愿景需要跨学科的深度协作，尤其在计算病理学与系统生物学的交叉领域，亟需建立统一的多模态数据表示方法和跨尺度特征对齐机制。只有当机器学习模型真正内化肿瘤生物学的第一性原理时，才能实现从“黑箱预测”到“透明推理”的质变，最终推动结直肠癌诊疗进入真正的智能时代[17]。

## 基金项目

国家自然科学基金(编号：8236030102)，云南省科技厅昆明医科大学应用基础研究联合专项基金(编号：202301AY070001-025)，兴滇英才支持计划(编号：XDYC-MY-2022-0100)。

## 参考文献


[1] Takamatsu, M., Yamamoto, N., Kawachi, H., Chino, A., Saito, S., Ueno, M., *et al*. (2019) Prediction of Early Colorectal Cancer Metastasis by Machine Learning Using Digital Slide Images. *Computer Methods and Programs in Biomedicine*, **178**, 155-161. https://doi.org/10.1016/j.cmpb.2019.06.022

[2] Hatzidaki, E., Iliopoulos, A. and Papasotiriou, I. (2021) A Novel Method for Colorectal Cancer Screening Based on Circulating Tumor Cells and Machine Learning. *Entropy*, **23**, Article No. 1248. https://doi.org/10.3390/e23101248

[3] 周龙妹, 王艳玲, 尹春英, 等. 机器学习算法预测模型在结直肠癌筛查中的应用进展[J]. 山东医药, 2023(35): 96-99.

[4] Yang, X., Yu, W., Yang, F. and Cai, X. (2023) Machine Learning Algorithms to Predict Atypical Metastasis of Colorectal Cancer Patients after Surgical Resection. *Frontiers in Surgery*, **9**, Article ID: 1049933. https://doi.org/10.3389/fsurg.2022.1049933

[5] Guo, Z., Zhang, Z., Liu, L., Zhao, Y., Liu, Z., Zhang, C., *et al*. (2024) Machine Learning for Predicting Liver and/or Lung Metastasis in Colorectal Cancer: A Retrospective Study Based on the SEER Database. *European Journal of Surgical Oncology*, **50**, Article ID: 108362. https://doi.org/10.1016/j.ejso.2024.108362

[6] Pang, L., Sun, Q., Wang, W., Song, M., Wu, Y., Shi, X., *et al*. (2025) A Novel Gene Signature for Predicting Outcome in Colorectal Cancer Patients Based on Tumor Cell-Endothelial Cell Interaction via Single-Cell Sequencing and Machine Learning. *Heliyon*, **11**, e42237. https://doi.org/10.1016/j.heliyon.2025.e42237

[7] Ting, W., Lu, Y.A., Ho, W., Cheewakriangkrai, C., Chang, H. and Lin, C. *et al*. (2020) Machine Learning in Prediction of Second Primary Cancer and Recurrence in Colorectal Cancer. *International Journal of Medical Sciences*, **17**, 280-291. https://doi.org/10.7150/ijms.37134

[8] Wang, Z., Shao, Y., Zhang, H., Lu, Y., Chen, Y., Shen, H., *et al*. (2023) Machine Learning-Based Glycolysis-Associated Molecular Classification Reveals Differences in Prognosis, TME, and Immunotherapy for Colorectal Cancer Patients. *Frontiers in Immunology*, **14**, Article ID: 1181985. https://doi.org/10.3389/fimmu.2023.1181985

[9] Burnett, B., Zhou, S., Brophy, S., Davies, P., Ellis, P., Kennedy, J., *et al*. (2023) Machine Learning in Colorectal Cancer Risk Prediction from Routinely Collected Data: A Review. *Diagnostics*, **13**, Article No. 301. https://doi.org/10.3390/diagnostics13020301

[10] Qin, L., Mao, J., Gao, M., Xie, J., Liang, Z. and Li, X. (2024) Machine Learning Models Can Predict Cancer-Associated Disseminated Intravascular Coagulation in Critically Ill Colorectal Cancer Patients. *Frontiers in Pharmacology*, **15**, Article ID: 1478342. https://doi.org/10.3389/fphar.2024.1478342

[11] Taguchi, N., Oda, S., Yokota, Y., Yamamura, S., Imuta, M., Tsuchigame, T., *et al*. (2019) CT Texture Analysis for the Prediction of KRAS Mutation Status in Colorectal Cancer via a Machine Learning Approach. *European Journal of Radiology*, **118**, 38-43. https://doi.org/10.1016/j.ejrad.2019.06.028

[12] Yu, Z.Y., *et al*. (2024) Machine Learning Application Identifies Plasma Markers for Proteinuria in Met-Astatic Colorectal Cancer Patients Treated with Bevacizumab. *Cancer Chemotherapy and Pharmacology*, **93**, 587-593.

[13] Liu, X., Shu, X., Zhou, Y. and Jiang, Y. (2024) Construction of a Risk Prediction Model for Postoperative Deep Vein Thrombosis in Colorectal Cancer Patients Based on Machine Learning Algorithms. *Frontiers in Oncology*, **14**, Article ID: 1499794. https://doi.org/10.3389/fonc.2024.1499794

[14] Ji, X.W., Wang, L.X., Luan, P.B., *et al*. (2025) The Impact of Dietary Fiber on Colorectal Cancer Patients Based on Machine Learning. *Frontiers in Nutrition*, **12**, Article ID: 1508562.

[15] Chang, C.-C. and Chen, Y.-C. (2020) Advanced Machine Learning in Prediction of Second Primary Cancer in Colorectal Cancer. In: *Studies in Health Technology and Informatics*, IOS Press, 1191-1192. https://doi.org/10.3233/shti200357

[16] Cai, L., Yang, D., Wang, R., Huang, H. and Shi, Y. (2024) Establishing and Clinically Validating a Machine Learning Model for Predicting Unplanned Reoperation Risk in Colorectal Cancer. *World Journal of Gastroenterology*, **30**, 2991-3004. https://doi.org/10.3748/wjg.v30.i23.2991

[17] Rhanoui, M., Mikram, M., Amazian, K., Ait-Abderrahim, A., Yousfi, S. and Toughrai, I. (2024) Multimodal Machine Learning for Predicting Post-Surgery Quality of Life in Colorectal Cancer Patients. *Journal of Imaging*, **10**, Article No. 297. https://doi.org/10.3390/jimaging10120297