

DOI:10.13288/j.11-2166/r.2026.07.012

基于多模态大模型的多中心中医肛肠优势病种专病数据平台的构建与应用

朱雨欣¹, 赵莉萍², 陆家发³, 朱慧婷¹, 杨霞¹, 杜蕾¹, 丁康[✉]

1. 南京中医药大学附属南京市中医院, 江苏省南京市秦淮区大明路157号, 210022; 2. 江苏省中医药管理局; 3. 江苏省卫生健康信息中心

[摘要] 构建基于多模态大模型的混合痔中医专病数据集平台并对其进行初步应用验证。平台采用 Star-Rocks 建立四级数据仓库体系, 实现多源异构数据的汇聚、清洗与标准化; 以 DeepSeek-R1-Distill-Qwen-7B 大模型为基座, 通过低秩自适应 (LoRA) 技术进行领域微调, 结合 LLaMA-3.3 自然语言处理与思维链技术, 实现对非结构化中医病历的智能解析与结构化提取, 准确识别症状、证候等 6 大类、28 子类实体, 微调后模型 F1 值达 93.8%。平台建立了包含超 5 万份病历的高质量专病数据集, 并已应用于一项涉及 17 831 例患者的真实世界研究, 初步验证了中医传承手术的疗效。

[关键词] 痔病; 多模态大模型; 专病数据集; 中医优势病种; 真实世界研究

混合痔作为肛肠科临床发病率最高的良性疾病之一, 其诊疗过程涉及复杂的解剖结构辨识、病理生理改变评估及与治疗相应的护理方式。混合痔为中医优势病种之一, 基于独特的中医辨证施治体系, 众多临床医生在长期临床诊疗过程中总结出丰富的中医特色方案, 包括硬化剂注射治疗^[1]、特有的手术方式以及围手术期的中药熏洗^[2]等。这些宝贵的临床经验与数据沉淀在各级医疗机构的电子病历 (electronic medical records, EMR) 系统中, 由于数据标准不一、系统架构封闭等原因, 最终导致“数据孤岛”的形成。

现有的医疗大数据平台研究多集中于文本结构化程度较高的西医领域, 针对中医专病的医疗大数据平台较少, 建设中医优势病种专病数据集平台面临极为特殊的挑战。一方面, 中医临床数据具有高度的多模态特征, 既包含基于解剖学的结构化指标, 又包含大量描述性的非结构化文本, 这两类数据在语义空间上存在巨大的异构性。另一方面, 中

医师对中医术语的描述具有个人理解, 导致不同医师对同一证候的描述可能存在显著差异, 由此带来的数据模糊性与多样性使得跨中心的数据融合异常困难。此外, 医疗数据的隐私敏感性限制了数据的物理汇聚, 传统的集中式机器学习范式在多中心协作场景下难以实施。基于此, 本研究联合南京市中医院、无锡市中医医院、徐州市中医院、张家港市中医医院、常熟市中医院、苏州市中医医院、江阴市中医院、泰州市中医院等 8 家三甲医院, 构建一个基于多模态大模型的多中心混合痔专病数据集平台。该平台拟解决多中心数据治理问题, 实现中医临床文本与结构化数据的深层语义对齐, 从而为混合痔的精准诊疗、并发症预测及中医科研提供智能化的数据底座。

1 优势病种专病数据平台框架的设计

为系统推进中医优势病种多中心临床研究的智能化进程并提升诊疗辅助水平, 本研究设计并构建了一套基于大数据与现代人工智能 (AI) 框架的医学大模型诊疗数据平台。该平台的核心目标在于整合分散的多源异构中医临床数据, 通过严谨的数据治理与先进的 AI 建模技术, 为高质量临床研究提供坚实基础, 并赋能精准、高效的智能诊疗应用。

基金项目: 江苏省中医肛肠疾病临床医学创新中心重点项目 (苏中医科教函〔2023〕1 号); 南京大学中国医院改革发展研究院全国范围内拟立项课题 (NDYG029, NDYG025); 江苏省医院协会医院管理创新课题 (JSYGY-3-2023-511)

✉ 通讯作者: 51422503@qq.com

平台数据架构采用高性能分布式分析型数据库 StarRocks 进行支撑,并建立层次化、规范化的四级数据仓库体系,即原始数据首先汇聚于贴源层(ODS),记录来自各合作医疗中心的初始诊疗信息;随后进入明细层(DWD),在此层面对数据进行深度清洗(处理缺失值与异常值)、标准化(统一术语编码,如症状、证候、方剂名称)及归一化处理(如剂量单位转换、时间格式统一),生成高质量、原子粒度的结构化基础数据;之后进入专病数据层(DWS),依据特定中医优势病种(如消渴病、中风病)的诊疗范式,对DWD层数据进行主题建模与维度汇总,形成服务于病种深度分析的高维宽表数据集;最后临床研究数据层(ADS)直接面向研究需求,提供经过高度聚合与指标计算的定制化数据集(如患者队列基线特征表、疗效评价指标表、生存分析表),以满足复杂统计分析与可视化呈现的要求。

平台的智能引擎核心选用具备强大中文语义理解与生成能力的开源大语言模型 DeepSeek-R1-Distill-Qwen-7B 作为基座。为使其深度适配中医专病领域知识,本研究采用了高效的低秩自适应(LoRA)微调技术。此技术仅需优化模型内部少量低秩矩阵参数,即可实现模型能力的精准迁移。至关重要是,微调所依赖的训练数据集严格来源于经过前述DWD层深度治理与DWS层专病构建的高质量、结构化/半结构化专病数据,可确保模型学习到的知识符合中医临床实践规范与逻辑。基于此微调模型,平台实现了两大核心智能应用功能:其一为病历信息智能结构化提取,能够自动识别非结构化病历文本中的关键临床实体及其相互关系,显著提升临床数据利用效率;其二为辅助诊疗决策支持,模型能够根据输入的患者信息,结合其学习到的专病知识图谱雏形与辨证逻辑,生成证候倾向性分析、鉴别诊断建议及参考方剂推荐,为临床医师提供决策参考。

为进一步提升辅助诊疗的准确性与中医理论契合度,克服纯数据驱动模型在复杂辨证中的局限性,平台基于治理后的高质量专病数据集,系统构建了“方剂-证候-适应症状/体征”等多维关联的中医专病知识图谱。该图谱明确定义了核心知识实体,并利用图计算技术深度挖掘实体间丰富的语义关系。此知识图谱的关键作用在于对AI模型的输出进行逻辑校验与语义增强,提升辅助诊疗建议的临床合理性与准确性。优势病种数据平台框架设计

图详见附图1(请扫描本文二维码获取)。

2 混合痔优势病种专病数据集的构建

2.1 专病数据集的质控

为了实现8家医院数据的有效融合,首要任务是建立一套科学、规范且具有中医特色的混合痔核心数据元(common data elements, CDE)标准。这不仅是数据清洗的基础,也是后续知识图谱构建与模型训练的基石。本数据库建设参考了《中国痔病诊疗指南(2020)》^[3]、欧洲结直肠外科学会发布的《痔疮疾病指南》^[4]等国内外医疗行业的临床实践指南,遵循国家卫生健康委员会制定的行业数据标准、国际相关术语标准等多项权威数据标准,由主管部门制定政策对专病数据集进行质控,达到了混合痔病历书写标准化,最终实现临床数据的统一与规范。

2.2 基础数据信息的纳入

数据库纳入的数据主要包含两个方面。一方面,纳入江苏省内所有医疗机构门急诊诊断包含“混合痔”、住院系统病案首页中入院诊断或出院诊断包含“混合痔”的患者的全部诊疗相关信息。主要包括:门诊病历、复诊病历、门诊处方、门诊检验、门诊检查、门诊摘要、门诊处置、门诊收费、病案首页、入院记录、病程记录、手术记录、出院小结、检验报告、内镜报告、病理报告、护理记录、随访记录等^[5]。另一方面,纳入中医经典、西医研究、中西医药物数据等多维度医学知识,提供丰富的结构化和非结构化数据源。

2.3 专病结构化大模型的微调

2.3.1 大语言模型的优化 大语言模型为医疗文本结构化提供了新的技术路径^[6]。相较于传统规则或机器学习方法,大语言模型能通过语义理解处理术语变异性问题,并捕捉跨症状的隐含关联。DeepSeek-R1-Distill-Qwen-7B 作为专为推理任务优化的中文大模型,其知识蒸馏架构和7B参数量级使其在资源受限场景中具有显著优势,但直接将通用大模型应用于专病病历结构化存在领域适配不足问题。为此,本研究参考基于LoRA的领域微调方案^[7],在保留原模型通用知识的前提下,注入专病诊疗知识,实现病历的精准结构化。

2.3.2 病历高质量全结构化数据训练集的构建

收集8家医院肛肠科2019—2024年混合痔病历,经脱敏处理获得10 000份文本,覆盖各年龄段、病程阶段及并发症类型。由肛肠科医师团队按标准标

注,定义6大类(现病史、中医证型、专科检查、既往史、治疗方式、实验室检查)、28子类实体(如症状、辨证、治法等),最终生成3000条训练集样本。

2.3.3 LoRA 的微调技术方案 LoRA 通过低秩分解矩阵实现参数高效微调,其数学表达为: $W' = W + \Delta W = W + A \cdot B$ 。

其中原始权重矩阵 $W \in \mathbb{R}^{(d \times k)}$ 冻结不变,仅训练低秩矩阵 $A \in \mathbb{R}^{(d \times r)}$ 与 $B \in \mathbb{R}^{(r \times k)}$ ($r < d, k$)。本方案采用以下关键配置:秩维度, $r = 64$ (平衡性能与效率);学习率, $2e^{-5}$;适配层, Q/K/V 矩阵+MLP 层。

2.3.4 模型的评估 评估模型采用实体级评估指标,包括准确率、召回率、F1 值。其中准确率为完全匹配的实体占比,召回率为识别出的实体占标注实体的比例,F1 值为准确率与召回率的调和平均值。

2.3.5 后结构化的训练 AI 中医病历高质量全结构化数据集构建过程中需对医师撰写的文本型非结构化数据进行事后处理,借助自然语言处理^[8]技术分析文本隐含的语义和上下文结构关系,从中发现并提取结构化所需的知识单元,这一过程称为后结构化。中医病历中的四诊信息和病程记录包含大量非结构化文本,是挖掘中医辨证规律的宝库。借助 LLaMA-3.3 自然语言处理功能对中医诊疗文本中的非结构化数据开展智能解析,利用正则表达式将病历文本分割为“主诉”“现病史”“既往史”“中医诊断”“西医诊断”等语义块,有助于模型区分当前症状与既往症状,避免时序混淆^[9]。针对中医文本古文与现代文混杂、专业术语密集的特点,采用挂载中医专业词典(包含方剂名、穴位名、证候名)的分词器进行初步分词,并利用正则表达式去除无意义的标点符号、特殊字符及系统自动生成的冗余模板文本。对于“专科检查”等半结构化文本,利用正则表达式提取关键信息,将其转换为结构化的 Key-Value 对,为后续结构化的构建提供精准的属性输入。

在此过程中,整合中医真实世界病历数据、中医药教材数据、江苏省中医优势病种诊疗经验共识数据等多源信息,开展大样本量 AI 训练,并对训练结果集进行开发与分类。随后随机抽取人工标注与 AI 标注的结果进行正确率审核,通过持续更新迭代不断提升 AI 识别的正确率。

2.3.6 基于不确定性的分层抽样与验证 传统随机采样对所有样本一视同仁,导致人工大量审核模

型已有 99% 把握的简单样本,造成算力与人力浪费。本研究采用基于模型置信度的分层采样策略。模型输出结果时,不仅生成标注文本,同时输出其对该预测的置信度分数,以量化模型的不确定性。根据不确定性程度将样本划分为不同层级,设定非均匀的采样比例^[10]。

为了最大化人工审核的收益,对不同层级采用差异化的抽样比例:对于层级 A 即置信度分数 > 0.9 的高置信度文本,采取抽取 5% 的低频抽样方式;对于层级 B 即 $0.6 < \text{置信度分数} < 0.9$ 的模糊区间文本,采取抽取 50% 的高频抽样方式;对于层级 C 即置信度分数 < 0.6 的低置信度文本,采取抽取 100% 的高频抽样方式。然后由两名中医肛肠医师独立对抽样病历进行结构化标注。最后由第三位资深主任医师介入,针对标注不一致的实体进行最终裁定。

由于采用了非均匀采样,困难样本被过度采样,直接计算准确率会低于模型在真实全量数据中的表现。因此,在报告最终模型效能时,必须根据各层级在总样本中的占比进行加权还原,随后审核 AI 标注结果的正确率,通过持续更新、迭代不断提升 AI 识别的正确率。

2.4 大数据集平台的管理

数据安全保障使用统一的认证和授权平台来进行,系统通过用户权限限制用户对系统级资源的访问包括菜单、数据、注册表项、数据库对象、日志文件等。对项目系统中的各类数据进行安全管理;对所有的数据资产进行识别、进入和统一管理,依据数据分级分类规则进行管理,从而确定相应的数据保护规则;根据不同类别和级别数据的保护规则,对数据进行加密或脱敏处理。

数据更新是针对已挂接的多中心数据资源,支持资源更新状态监听功能,周期性监测资源更新情况,评估资源是否按期更新,支持在线对文件、库表类型的资源进行更新,可点击查看详细更新记录。通过记录数据的版本信息,可以追踪数据的变更历史,确保数据的完整性和准确性。同时,版本控制还可以帮助使用者了解数据的最新状态,避免使用过时的数据。

3 中医优势病种多模态大模型专病数据集平台的构建及应用

基于以上数据集的设计及构建,最终构建了包含四级数据仓库、多模态智能引擎及中医知识图谱

的“三位一体”智能化多中心诊疗数据平台，并形成一套可推广的混合痔专病数据集治理范式。建成的多中心中医优势病种诊疗数据库覆盖并整合江苏省内几乎所有医疗机构混合痔患者全诊疗数据（门急诊、住院、随访等全流程信息包括通用模型 11 个、专病模型 25 个）。数据类型包括临床病历（门诊/住院记录、手术报告等）、检验/影像/病理报告、中医经典文献与药物数据库。数据规模以混合痔为试点，积累超过 5 万份脱敏病历文本，形成高质量标注数据集。目前实现了混合痔、社区获得性肺炎、溃疡性结肠炎、桡骨远端骨折等 4 个优势病种诊疗经验的推广应用、21 家试点医疗机构病历改造升级和中医诊疗数据采集上传。

3.1 中医优势病种混合痔专病数据集大模型结构化引擎构建及模型评价

构建的混合痔专病大模型结构化引擎，基于 DeepSeek-R1-Distill-Qwen-7B 模型，采用 LoRA 微调技术（秩 $r=64$ ），注入混合痔专科知识；结合 LLaMA-3.3 与 DeepSeek-R1 思维链技术，实现中医文本的深度语义解析并进行性能突破，实现自动解析非结构化病历，提取 6 大类、28 子类中医实体。一方面生成符合科研要求的结构化数据集功能，为临床研究提供高质量结构化数据集；以混合痔为模板，验证了从数据采集→标准化→AI 结构化→专病库构建的全链路可行性，为后续扩展至其他优势病种（如糖尿病、慢性肾病）提供技术模板。另一方面，大模型推理系统负责处理大规模数据的推理任务，系统实时接收和处理医疗数据，并立即进行模型推理，以快速生成诊断或分析结果，辅助临床医生进行医疗决策。

DeepSeek-R1-Distill-Qwen-7B 模型准确率为 58.3%，召回率为 52.7%，F1 值为 55.3%，微调后模型准确率、召回率、F1 值分别为 94.5%、93.1%、93.8%。可见，LoRA 微调在准确率和 F1 值方面均有明显提升。AI 标注结果与人工标注结果对比结果详见表 1。最终中医优势病种混合痔专病数据集大模型结构化正确率 > 90%。

3.2 基于专病数据集构建完成混合痔的人群数据集队列及质控方案

完成构建了以混合痔为典型中医优势病种的全周期专病队列。该队列完整覆盖了从门诊到住院的诊疗全过程，系统整合了疾病分型、共病模式、亚组特征、高危因素及智能随访等多维度数据集。完成内涵质控评价行业专属模型，并围绕混合痔门

表 1 人工智能标注抽样病历结果与人工标注
抽样病历结果对比

Table 1 Comparison of accuracy between AI-annotated and manually annotated sampled medical records

指标	合计标签数	正确标签数	错误标签数	正确率/%
门诊既往史	268	267	1	99.63
门诊现病史	129	112	17	86.82
门诊专科检查	135	119	16	88.15
住院既往史	112	112	0	100.00
住院现病史	1203	1085	118	90.19
住院专科检查	689	607	82	88.10
住院出院情况	1365	1231	134	90.18
总计	3901	3533	368	90.57

诊、住院诊疗全过程，从中医药经验传承运用视角，利用大模型对全省试点医院混合痔诊疗过程内涵进行质控分析，建立标准化的数据利用机制，有效支持真实世界研究，促进科研院所、药企及行业第三方实现数据要素的规范化流转与商业化应用。

3.3 多中心中医优势病种专病数据集成果初步应用及验证

目前已基于混合痔多中心中医优势病种专病数据集开展了一项真实世界多中心、前瞻性、队列研究。该研究利用诊疗数据库挖掘优势病种诊疗方案中中医传承手术治疗混合痔的有效性、安全性。其中中医传承手术包括内痔套扎术、分段齿形结扎术等，接受中医传承手术/分段齿形结扎术治疗混合痔的患者为暴露组，不限定是否为首次使用，不限定疾病分型/适应证，不限定合并疾病、合并用药或治疗。所有没有使用中医优势病种诊疗方案中中医传承手术治疗而接受西医手术的受试者为非暴露组。初步研究纳入住院手术患者共 17 831 例（暴露组 6185 例，非暴露组 11 646 例）。出院时暴露组总体治愈率 92.7%，非暴露组总体治愈率为 90.0%，两组疗效差异无统计学意义 [$OR=0.87$, 95% CI (0.74, 1.06), $P=0.183$]。术后非暴露组未治愈患者 30 例，暴露组未治愈患者 14 例，两组未愈状态有统计学差异 [$OR=1.51$, 95% CI (1.04, 2.20), $P=0.032$]。随访及远期疗效观察将在后续研究中进一步开展。

4 结语

从临床疾病管理^[11]、医技辅助^[12]到药物开发^[13]，AI 已经深入临床医学的各过程。多中心中医优势病种专病数据集平台利用大规模多模态语言模型，针对医学领域进行微调训练，构建医学领域

的大语言模型,实现医学文本的智能理解和生成,解决了医疗系统数据采集分析面临的挑战^[14],最终形成具有中医优势病种特有的中医药大模型平台。该平台通过大模型归纳整合数据,挖掘疾病信息,促进混合痔诊疗数据向信息化、数字化、智能化转换,充分利用新技术新方法,构建新的中医药临床研究范式^[15]。平台整合了江苏省中医药混合痔专病病历的后结构化知识中心,创新性地采用 LLaMA-3.3 的自然语言处理能力实现中医诊疗数据文本的智能解析,通过基于思维链的 DeepSeek-R1 深度思考技术,构建涵盖混合痔这一中医优势病种的动态实体关系图谱,为混合痔中医药真实世界研究提供了智能化技术支撑。将 AI 大模型和大数据平台充分引入中医学是当今中医药发展的趋势,也是促进中医药走向世界的关键。随着技术等相关领域的进步,中医各优势病种医学大模型诊疗数据平台也会随之发展,以进一步促进中医药在新时代焕发新光彩。

致谢

江苏省中医优势病种诊疗大数据平台支持。

参考文献

- [1]李冠路,高记华,徐丹阳,等. 中医硬化注射疗法治疗肛肠疾病现状与进展[J]. 中华中医药杂志,2022,37(6): 3351-3355.
- [2]杨会举,刘佃温. 复方黄柏液涂剂治疗混合痔临床应用专家共识(2025 年)[J]. 中国新药杂志,2025,34(19): 2122-2128.
- [3]中国中西医结合学会大肠肛门病专业委员会. 中国痔病诊疗指南(2020)[J]. 结直肠肛门外科,2020,26(5): 519-533.
- [4]VAN TOL RR, KLEIJNEN J, WATSON AJM, et al. European Society of ColoProctology: guideline for haemorrhoidal disease[J]. Colorectal Dis,2020,22(6):650-662.
- [5]KONG X, PENG R, DAI H, et al. Disease-specific data processing: an intelligent digital platform for diabetes based on model prediction and data analysis utilizing big data technology [J]. Front Public Health, 2022, 10: 1053269. doi: 10.3389/fpubh.2022.1053269.
- [6]OMIYE JA, GUI H, REZAEI SJ, et al. Large language models in medicine: the potentials and pitfalls : a narrative review[J]. Ann Intern Med,2024,177(2):210-220.
- [7]JIA J, NISHI H. A flexible two-stage anonymization framework for narrative medical records adapting to various language models [J]. Comput Biol Med, 2025, 195: 110624. doi: 10.1016/j.compbimed.2025.110624.
- [8]BÜTTNER M, LESER U, SCHNEIDER L, et al. Natural language processing: chances and challenges in dentistry [J]. J Dent, 2024, 141: 104796. doi: 10.1016/j.jdent.2023.104796.
- [9]JIA Q, ZHANG D, YANG S, et al. Traditional Chinese medicine symptom normalization approach leveraging hierarchical semantic information and text matching with attention mechanism [J]. J Biomed Inform, 2021, 116: 103718. doi:org/10.1016/j.jbi.2021.103718.
- [10]YANG X, LI S, SHAO Q, et al. Uncertainty-guided man-machine integrated patient-specific quality assurance [J]. Radiother Oncol,2022,173:1-9. doi: 10.1016/j.radonc.2022.05.016.
- [11]DUBEY S, CHAN A, ADEBAJO AO, et al. Artificial intelligence and machine learning in rheumatology [J]. Rheumatology (Oxford),2024,63(8):2040-2041.
- [12]HANNA MG, PANTANOWITZ L, DASH R, et al. Future of artificial intelligence-machine learning trends in pathology and medicine [J]. Mod Pathol, 2025, 38(4): 100705.
- [13]GHOLAP AD, UDDIN MJ, FAIYAZUDDIN M, et al. Advances in artificial intelligence for drug delivery and development: a comprehensive review [J]. Comput Biol Med, 2024,178:108702. doi:org/10.1016/j.compbimed.2024.108702.
- [14]SWEENEY SM, HAMADEH HK, ABRAMS N, et al. Challenges to using big data in cancer [J]. Cancer Res, 2023,83(8):1175-1182.
- [15]中华医学会临床流行病学和循证医学分会中医学组. 新时代中医药临床研究方法论专家共识[J]. 协和医学杂志,2022,13(5):783-788.

Construction and Application of a Multicenter Traditional Chinese Medicine Proctology Disease Data Platform Based on Multimodal Large Models

ZHU Yuxin¹, ZHAO Liping², LU Jiafa³, ZHU Huiting¹, YANG Xia¹, DU Lei¹, DING Kang¹

1. Nanjing Hospital of Chinese Medicine affiliated to Nanjing University of Chinese Medicine, Nanjing, 210022; 2. Traditional Chinese Medicine Administration, Jiangsu Commission of Health; 3. Jiangsu Provincial Health Information Center

ABSTRACT This paper has constructed a traditional Chinese medicine (TCM) specialized disease dataset platform for mixed hemorrhoids based on a multimodal large model, and the preliminary application has been validated. The platform uses StarRocks to establish a four-level data warehouse system, enabling the aggregation, cleaning, and standardization of multi-source heterogeneous data. Using DeepSeek-R1-Distill-Qwen-7B as the base model, domain fine-tuning is performed through low-rank adaptation (LoRA) technology. Combined with LLaMA-3.3 natural language processing and reasoning chain techniques, the platform enables intelligent parsing and structured extraction of unstructured TCM medical records. It accurately identifies six major categories and 28 subcategories of entities, including symptoms and syndromes, with a fine-tuned model F1 score of 93.8%. The platform has established a high-quality specialized disease dataset containing more than 50,000 medical records and has been applied in a real-world study involving 17,831 patients, preliminarily verifying the efficacy of TCM heritage surgery.

Keywords hemorrhoids; multimodal large model; specialized disease dataset; dominant diseases of TCM; real-world study

(收稿日期: 2025-09-17; 修回日期: 2026-01-16)

[编辑: 贾维娜]

(上接第 769 页)

Construction of An Automated Segmentation Visual Foundation Model for Pathological Images of Hemorrhoids and Its Application in Traditional Chinese Medicine Clinical Syndrome Analysis

ZHANG Shijie¹, ZHANG Ao¹, WANG Kang¹, KANG Bin², YU Xiaofan², FENG Xujing¹, CAO Jinyu¹, HUANG Wenzhen¹, DING Kang¹

1. Nanjing Hospital of Chinese Medicine Affiliated to Nanjing University of Chinese Medicine, Nanjing, 210022; 2. School of Internet of Things, Nanjing University of Posts and Telecommunications

ABSTRACT This paper proposes a two-stage method integrating visual foundation models (VFM) and diffusion models. The segment anything model (SAM) as VFM is combined with the SegRefiner diffusion model to construct the SAM-SegRefiner framework for automated segmentation of edema, inflammation, and thrombus regions in histopathological images of hemorrhoidal tissue, providing a reproducible technical tool for the objective quantification of pathological morphology and its application in traditional Chinese medicine (TCM) syndrome research. Trained and validated on multi-center retrospective data, the SAM-SegRefiner model achieved an average pixel accuracy of 95.32% and a mean intersection over union (mIoU) of 66.81% on an independent test set, significantly outperforming comparative models such as U-Net, MixU-Net, and SAM-Med2D, and also demonstrating robust cross-center generalization capability. Furthermore, by correlating the quantitatively segmented results from the model with the patients' TCM syndrome types, the potential associations between pathomorphological features and TCM syndrome differentiation have been explored. The analysis revealed no statistically significant differences in the degree of inflammatory infiltration and thrombus formation among different syndrome types, suggesting a complex relationship between local pathological changes and systemic syndrome manifestations.

Keywords hemorrhoids; pathological diagnosis; visual foundation models; segment anything model; traditional Chinese medicine syndrome

(收稿日期: 2025-10-30; 修回日期: 2026-01-20)

[编辑: 贾维娜]