Advances in Clinical Medicine 临床医学进展, 2026, 16(4), 1924-1934
Published Online April 2026 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2026.1641434

深度学习辅助痰涂片抗酸杆菌分割模型构建与多中心验证

徐小斐[1]，孙有湘[2]，王宏伟[2]，李　静[2]，张齐波[3]，王　清[2*]，高绪栋[4]，李培静[4]

[1]青岛大学医学院，山东 青岛
[2]青岛大学附属医院检验科，山东 青岛
[3]青岛市公共卫生临床中心(高新院区)检验科，山东 青岛
[4]青岛海泰新光科技股份有限公司，山东 青岛

收稿日期：2026年3月8日；录用日期：2026年4月2日；发布日期：2026年4月9日

摘　要

目的：构建并验证一种面向痰涂片抗酸杆菌(Acid-Fast Bacilli, AFB)像素级识别的深度学习分割模型，评价其在多中心场景下的泛化能力、稳定性与临床可应用性。方法：基于4家医疗中心319例患者数据，纳入5647张含目标像素级标注图像块，采用分阶段消融策略开展模型构建与优化。在统一训练框架下，依次比较多中心训练策略、负样本混入比例、编码器容量、注意力机制和输入分辨率；通过同中心测试、多中心验证、5折交叉验证及留一中心交叉验证(LOCO)进行综合评估。结果：最优配置为U-Net + ResNet50 + scSE + DiceBCE + Strong增强 + 512 × 512，测试集Dice为0.8760，IoU为0.7794，多中心Dice为0.8579。5折交叉验证Dice为0.8741 ± 0.0019，LOCO Dice为0.8591 ± 0.0124，泛化差距为1.50%。数据策略显示，简单多中心混合训练未形成稳定增益；负样本按1:1~1:3比例直接混入会降低泛化性能。单张图像推理约0.18 s。结论：在多中心痰涂片AFB分割任务中，以系统消融驱动的模型构建路径可同时获得较高精度、较低泛化差距和可接受推理效率，可为结核病实验室镜检流程的标准化与智能化提供技术支持。

关键词

抗酸杆菌，痰涂片，语义分割，多中心验证，深度学习

Deep Learning-Assisted Construction and Multicenter Validation of an AFB Segmentation Model on Sputum Smears

Xiaofei Xu[1], Youxiang Sun[2], Hongwei Wang[2], Jing Li[2], Qibo Zhang[3], Qing Wang[2*], Xudong Gao[4], Peijing Li[4]

*通讯作者。

文章引用: 徐小斐, 孙有湘, 王宏伟, 李静, 张齐波, 王清, 高绪栋, 李培静. 深度学习辅助痰涂片抗酸杆菌分割模型构建与多中心验证[J]. 临床医学进展, 2026, 16(4): 1924-1934. DOI: 10.12677/acm.2026.1641434

[1]School of Medicine, Qingdao University, Qingdao Shandong
[2]Department of Clinical Laboratory, The Affiliated Hospital of Qingdao University, Qingdao Shandong
[3]Department of Clinical Laboratory, Qingdao Public Health Clinical Center (High-Tech Zone), Qingdao Shandong
[4]Qingdao Novel Beam Technology Co., Ltd., Qingdao Shandong

Received: March 8, 2026; accepted: April 2, 2026; published: April 9, 2026

Abstract

Objective: To develop and validate a deep learning–based segmentation model for pixel-level recognition of Acid-Fast Bacilli (AFB) in sputum smear images, and to evaluate its generalization ability, robustness, and clinical applicability across multi-center settings. Methods: Based on data from 319 patients across four medical centers, a total of 5,647 annotated image patches containing target bacilli were collected. A staged ablation strategy was adopted for model construction and optimization. Within a unified training framework, we systematically compared multi-center training strategies, the proportion of negative sample mixing, encoder capacity, attention mechanisms, and input resolution. Model performance was comprehensively evaluated through intra-center testing, multi-center validation, five-fold cross-validation, and Leave-One-Center-Out (LOCO) validation. Results: The optimal configuration (U-Net + ResNet50 + scSE + DiceBCE + Strong Augmentation, 512 × 512) achieved a Dice coefficient of 0.8760 and an IoU of 0.7794 on the test set. Multi-center Dice was 0.8579. Five-fold cross-validation yielded a Dice of 0.8741 ± 0.0019, while LOCO validation achieved 0.8591 ± 0.0124, with a performance drop of less than 1.50%. Data strategy analysis showed that simple multi-center data mixing did not yield stable improvements, and incorporating negative samples at a ratio of 1:1 - 1:3 reduced model performance. The inference time per image was approximately 0.18 s. Conclusion: In multi-center AFB sputum smear segmentation tasks, a systematically ablation-driven model development strategy can achieve high accuracy, minimal generalization loss, and acceptable inference efficiency, thereby providing technical support for the standardization and intelligent transformation of laboratory diagnostic workflows for tuberculosis.

Keywords

Acid-Fast Bacilli, Sputum Smear, Semantic Segmentation, Multicenter Validation, Deep Learning

1. 引言

结核病防控仍是公共卫生体系中的关键任务。痰涂片抗酸杆菌镜检因检测成本低、周转快、基层可及性高，在我国及其他中高负担地区持续承担初筛与随访功能[1]-[3]。随着检验量增加，人工镜检在效率、重复性和一致性方面的瓶颈日益突出。

传统人工阅片对人员经验依赖显著，不同检验人员对弱阳性和边界模糊区域的判读差异较大，中心间的流程与标准也难完全统一。面对高通量任务，阅片疲劳和主观波动会进一步放大漏检与误检风险[3] [4]。

近年来，显微图像数字化与深度学习融合推动了智能辅助阅片的发展。与“是否阳性”的图像级判别相比，像素级分割可以更细致地呈现病灶位置和形态边界，便于专家复核、过程追踪和质量控制，因此在

临床实验室场景具有更强解释性[5]-[7]。此外像素级的精准定位也能极大地节省算力提高检验的经济性。

但从算法到临床的关键障碍并不在于单中心最高分，而在于跨中心稳健性。不同机构在染色批次、样本制备、运输保存、扫描习惯和背景杂质构成方面均可能存在差异，模型在本地数据上的高分并不必然转化为外中心可用[8] [9]。

现有研究中，针对 AFB 分割的多中心系统性证据仍相对有限，尤其缺少对“数据策略–模型结构–泛化验证”联动关系的分层分析[10] [11]。因此，临床实践中仍难回答“该检验模型为何有效、在何种场景有效、推广风险在哪里”等核心问题。

基于上述背景，本研究围绕多中心痰涂片 AFB 分割任务，构建分阶段消融框架，系统评估数据组织方式、网络架构与验证方案对泛化性能的影响，并结合推理效率结果讨论其流程落地潜力。研究目标是形成可复用、可解释、可验证的建模路径，为后续临床转化提供方法学依据的同时，提供区域性建模分析证据(见图 1)。

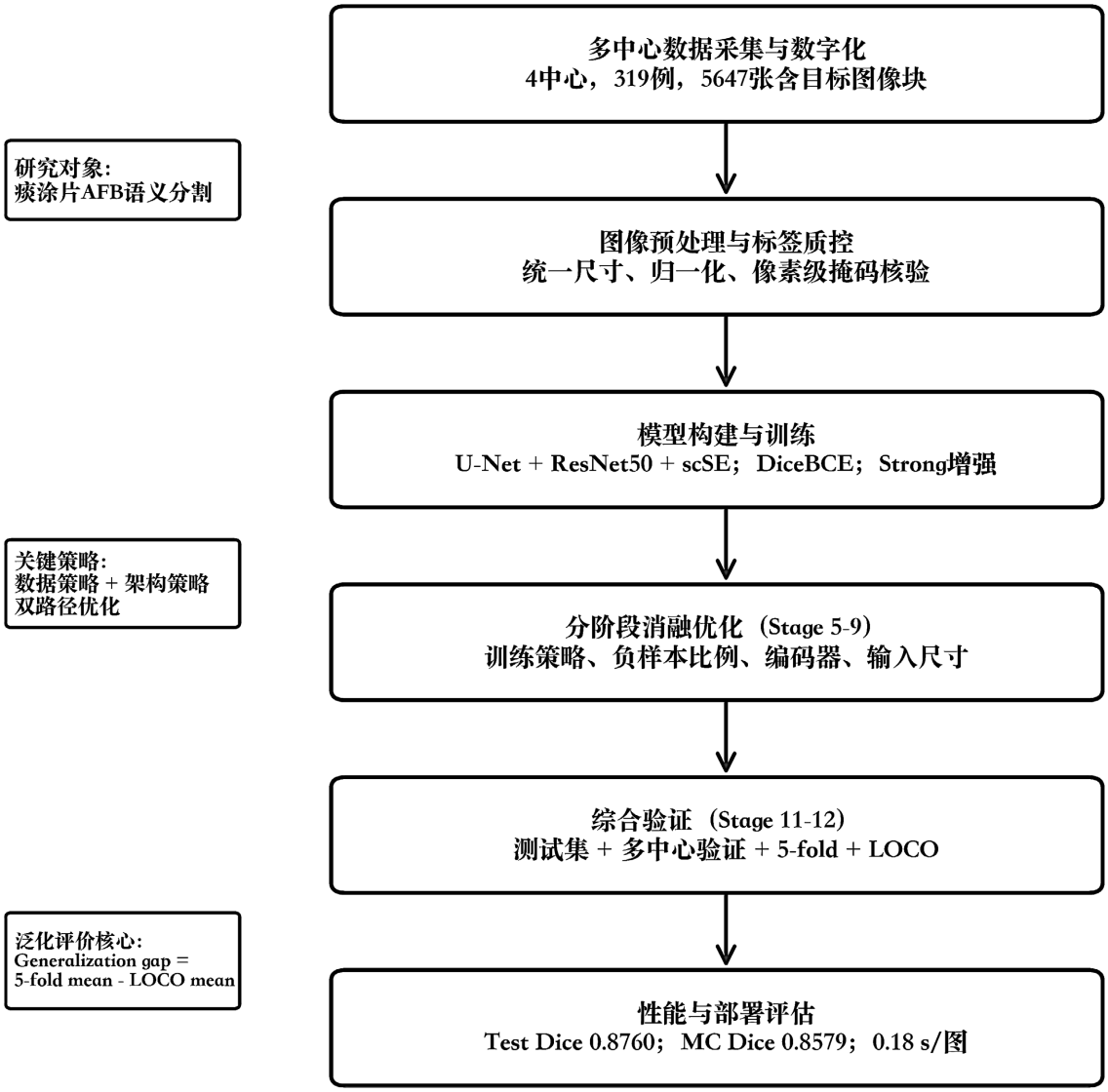


Figure 1. Study workflow diagram
图 1. 研究流程图

2. 资料与方法

2.1. 研究设计与数据来源

本研究为回顾性多中心方法学研究，数据来自 4 家医疗中心既有痰涂片数字显微图像库。通过海泰新光高通量抗酸杆菌数字扫描系统。图像采集模块由高分辨率的显微物镜、TUBE LENS、高清医用级显微数码相机及透射明场显微光源组成。利用高清医用级显微相机获取分辨率为 2448 * 2048 像素的痰涂片扫描图片。图像块提取流程如下：扫描系统对每张涂片进行全视野高通量扫描，内置预筛算法依据色度与形态特征自动标记候选视野；随后由经培训的检验人员对候选视野逐一复核，对确认含有抗酸杆菌的视野执行像素级标注。凡经人工确认标注且包含至少一个 AFB 前景区域的图像块即纳入“含目标”数据集。因此，本研究训练集以含目标图像块为主体，不含随机裁剪的纯阴性视野。这一构成特征是由标注资源约束和前期项目设计决定的，其对假阳性行为的潜在影响将在负样本消融实验(Exp-018~020)及局限性部分进一步讨论。共纳入 319 例患者，获得含目标像素级标注图像块 5647 张[3]。

数据划分采用双重验证逻辑：使用 Center 4 训练集 3203 张图像块进行训练，Center 4 测试集 427 张图像块用于同中心独立评估；Center 1-3 共 1376 张图像块用于跨中心验证，以模拟真实外部部署情境(见表 1)。

Table 1. Study cohort and dataset composition
表 1. 研究对象与数据集构成

变量	数值	说明
中心数量	4	Center 1~Center 4
患者总数	319	纳入病例总数
含目标图像块	5647	with_target 数据
训练集(中心 4)	3203	用于模型训练
测试集(中心 4)	427	同中心独立测试
多中心验证集(中心 1~3)	1376	跨中心泛化验证

该划分方式能够在“同域精度”与“跨域泛化”之间建立同时可观察的评估通道，减少仅凭单一测试集得出乐观结论的风险。

2.2. 图像标注与数据质控

本研究采用项目既有像素级掩码标签作为监督信号[5] [6]。建模前执行数据质控流程，包括图像–标签配对检查、分辨率一致性检查、异常值与损坏文件清理、重复样本去重和边界异常样本复核。

对于前景占比过低或标签区域突变样本，采用人工抽检方式进行二次确认，以降低标签噪声对训练稳定性的影响。通过强化质控，可减少模型将标注误差误判为“目标特征”的风险。

2.3. 预处理与增强策略

预处理包括尺寸统一、像素归一化及标签对齐。数据增强采用递进式消融：无增强、基础增强、中等增强和强增强。基础增强包含翻转和旋转；中等增强加入仿射与色彩扰动；强增强进一步加入形变操作，用于模拟跨中心成像波动[5] [8]。

增强策略实验完成后，Strong 增强作为后续主实验标准配置。该设计可在不改变标注语义的前提下

增加样本多样性，缓解模型对单一中心分布的过拟合。

2.4. 模型构建与消融分组

基础网络为 U-Net 分割框架[7]，基线配置为 ResNet50 编码器与 DiceBCE 损失。围绕该基线，分别开展数据策略消融和架构策略消融。

数据策略消融包括：单中心训练(Exp-015)、双中心混合训练(Exp-016)、四中心混合训练(Exp-017)；鉴于训练集以含目标图像块为主，模型在训练阶段缺少对纯阴性视野的充分暴露，可能在全视野推理中产生假阳性。为定量评估此风险并探索缓解策略，本研究设计了负样本混入实验：将不含 AFB 的纯背景图像块按正负样本 1:1 (Exp-018)、1:2 (Exp-019)、1:3 (Exp-020)比例混入训练集，以考察负样本规模对模型假阳性抑制能力与多中心泛化性能的联合影响。

架构策略消融包括：编码器容量比较(Exp-021、Exp-022) [5] [6]、注意力机制比较(含 Exp-025 scSE)、输入分辨率比较(Exp-027~Exp-030)。

在主要策略确定后，采用 5 折交叉验证和留一中心交叉验证开展稳定性与外推能力评估。

2.5. 评价指标与统计方法

主要指标为 Dice 与 IoU [5] [6]，分别反映分割重叠质量与交并比。辅助报告准确率和召回率，用于解释假阳性与假阴性行为。

为衔接像素级分割性能与临床样本级诊断需求，本研究进一步定义涂片级阳性判定规则：对每张涂片的全部图像块执行模型推理，提取各图像块分割掩码中的连通域并计数；当单张涂片中检出 AFB 连通域总数达到设定阈值时判定为模型阳性，否则为阴性。以临床病理报告为金标准，计算涂片级敏感性、特异性和总符合率，并在不同检出阈值(≥1、≥3、≥5)下分析其变化趋势，以评估像素级分割精度向样本级临床判断的传递效果。统计表示为均值 ± 标准差。泛化差距定义为 5-fold 平均 Dice 减去 LOCO 平均 Dice，即 Gap = Dice_5fold_mean-Dice_LOCO_mean。

为提高报告一致性，文中连续变量统一保留 4 位小数，百分比保留 2 位小数。

2.6. 软硬件环境与可复现性控制

实验在统一 GPU 服务器环境中完成，采用固定随机种子和统一训练调度策略。通过锁定训练轮次上限、早停规则和优化器参数，保证不同实验组之间仅改变目标变量。

模型推理效率在统一硬件与输入尺度下统计，避免由环境差异引入速度偏差。所有关键结果均来源于结构化实验记录文件，可用于后续审稿与复核。

2.7. 伦理与数据安全说明

本研究为回顾性图像数据分析，不涉及受试者额外临床干预。

纳入数据在建库与分析前均完成去标识化处理，仅保留建模所需字段。多中心数据在授权范围内进行汇总与使用，研究过程中采用受控服务器存储、分级权限访问与操作日志管理，确保数据安全与患者隐私保护。研究过程遵循医学伦理相关规范执行。

3. 结果

3.1. 多中心训练策略结果

单中心训练(Exp-015)获得 Test Dice 0.8749、Test IoU 0.7777、MC Dice 0.8510。双中心混合训练(Exp-016)降至 Test Dice 0.8706、MC Dice 0.8467。四中心混合训练(Exp-017)Test Dice 为 0.8655，MC Dice 为

0.8523 (详见表 2)。

从应用视角看，尽管 Exp-017 在 MC Dice 上略高于 Exp-015 (+0.0013)，但其同中心测试表现显著下降(−0.0094)，提示混合训练的收益并不稳定，且可能带来域间干扰。

Table 2. Results of multicenter training and negative sampling
表 2. 多中心训练与负样本策略结果

实验	策略	Test Dice	Test IoU	MC Dice
Exp-015	仅 Center4 训练	0.8749	0.7777	0.8510
Exp-016	Center1 + Center4 联合训练	0.8706	0.7709	0.8467
Exp-017	四中心联合训练	0.8655	0.7630	0.8523
Exp-018	负样本 1:1	0.8684	0.7676	0.8009
Exp-019	负样本 1:2	0.8640	0.7608	0.7923
Exp-020	负样本 1:3	0.8716	0.7725	0.8125

3.2. 负样本比例结果

负样本 1:1 混入(Exp-018)时 MC Dice 降至 0.8009；1:2 混入(Exp-019)时降至 0.7923；1:3 混入(Exp-020)时为 0.8125。三组策略均低于基线 0.8510。

相对基线，三组下降幅度分别为 5.01、5.87 和 3.85 个百分点。该结果表明，当前任务中“直接比例扩增负样本”并不等同于“提升背景判别能力”，反而可能削弱对微小目标的召回。进一步分析准确率与召回的变化模式发现，负样本混入后准确率呈上升趋势(基线 0.8637→Exp-018 为 0.8876→Exp-020 为 0.8981)，表明模型确实学习到更保守的前景预测策略，假阳性得到一定抑制；但 Recall 同步下降(基线 0.8863→Exp-018 为 0.8501→Exp-020 为 0.8468)，且 MC Dice 整体劣于基线。该结果表明，在前景极稀疏的 AFB 分割任务中，直接按比例混入负样本虽可降低假阳性，但会以牺牲跨中心召回能力为代价，二者之间存在权衡关系。这提示后续假阳性控制策略应探索课程式负样本引入或阈值后处理路径，而非简单比例扩增。

3.3. 架构优化结果

编码器比较中，ResNet101 (51.51M)对应 Test Dice 0.8688、MC Dice 0.8387；EfficientNet-B3 (13.16M)对应 Test Dice 0.8751、MC Dice 0.8544。后者在参数更少的情况下取得更优泛化表现。

注意力机制比较显示，scSE 配置(Exp-025)达到 Test Dice 0.8760、MC Dice 0.8579，为全流程最优泛化方案。

输入尺寸比较显示，256 × 256 下小目标细节损失明显；512 × 512 在精度与成本之间更平衡；640 × 640 未体现与成本增长相匹配的性能增益(见表 3)。

Table 3. Results of model construction and architecture optimization
表 3. 模型构建与架构优化结果

实验	策略	参数量(M)	Test Dice	Test IoU	MC Dice
Exp-021	ResNet101 编码器	51.51	0.8688	0.7680	0.8387
Exp-022	EfficientNet-B3 编码器	13.16	0.8751	0.7781	0.8544
Exp-025	ResNet50 + scSE	33.82	0.8760	0.7794	0.8579

续表

Exp-027	输入 256 × 256	32.52	0.8682	0.7673	0.8435
Exp-028	输入 384 × 384	32.52	0.8729	0.7746	0.8482
Exp-029	输入 512 × 512	32.52	0.8750	0.7778	0.8459
Exp-030	输入 640 × 640	32.52	0.8749	0.7777	0.8489

3.4. 多中心泛化验证结果

最优模型 5-fold Dice 为 0.8741 ± 0.0019，LOCO Dice 为 0.8591 ± 0.0124。前者标准差较小，说明模型训练稳定；后者在外中心保持较高水平，说明具有跨域适应潜力。

基于统一公式计算得到泛化差距 0.0150 (1.50%)，提示跨中心迁移性能衰减可控(见表 4)。

Table 4. Results of multicenter generalization validation
表 4. 多中心泛化验证结果

指标	结果
5-fold CV mean ± SD	0.8741 ± 0.0019
LOCO mean ± SD	0.8591 ± 0.0124
Generalization gap	0.0150 (1.50%)

3.5. 推理效率与流程可接入性结果

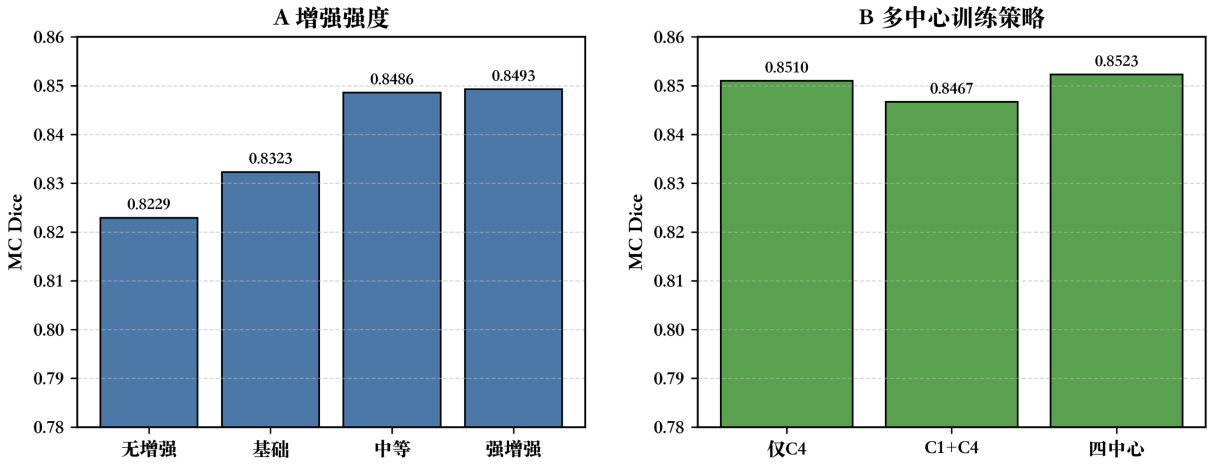
在 RTX 5090 环境下，512 × 512 输入时单张图像推理约 0.18 s；根据项目流程统计，整片处理约 12 s。该速度可支持实验室场景下的快速预标注与复核辅助。

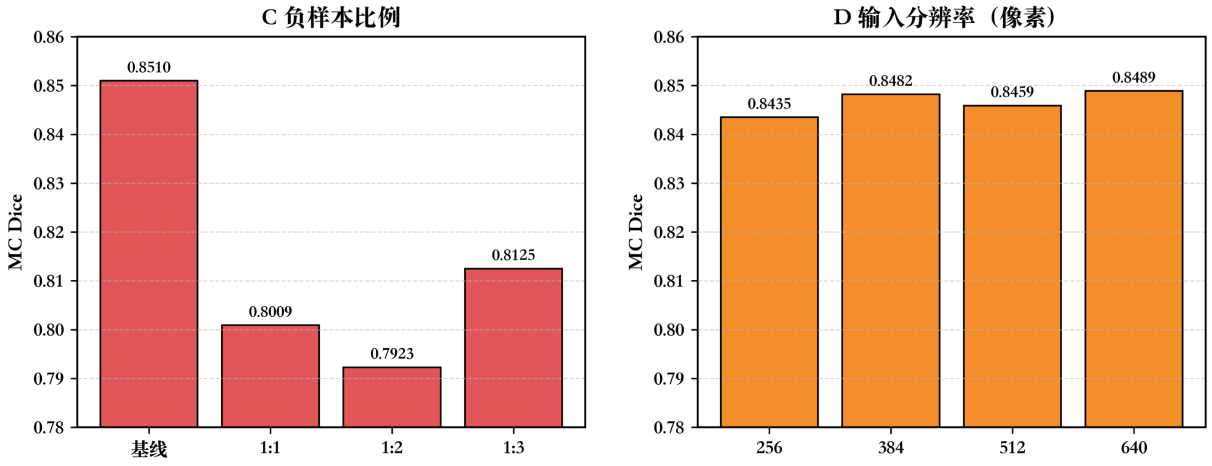
在高通量批次任务中，AI 可先完成候选区域筛选与风险提示，再由检验人员进行复核确认，从而减少重复性劳动并提升报告周转效率。

3.6. 错误模式分型观察

定性分析显示，假阳性主要出现在染色不均、背景杂质较多和局部对比度异常区域；假阴性主要集中于极细菌体碎片、低染色强度区域及密集聚团边界。

关键策略对多中心 Dice 的影响见图 2，这些模式提示后续优化应重点关注颜色归一化、困难样本再采样与边界敏感损失设计，以进一步提升复杂场景鲁棒性。四中心典型分割可视化结果见图 3。





注：A为增强消融；B-C为Stage 5-6结果；D为输入分辨率消融。

Figure 2. Impact of key strategies on multicenter Dice
图 2. 关键策略对多中心 Dice 影响

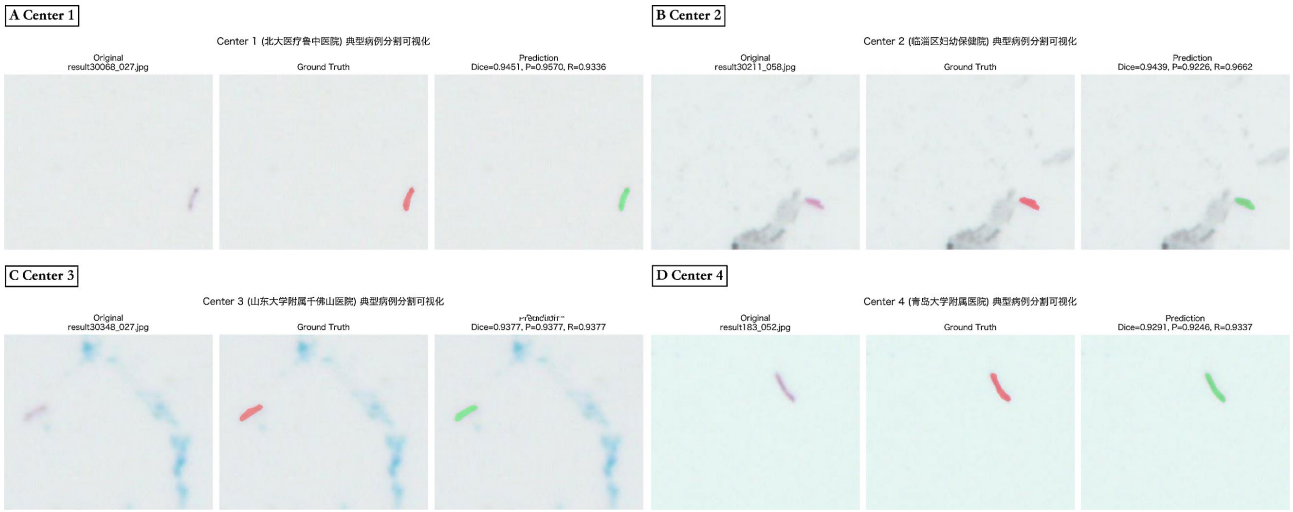


Figure 3. Visualization of typical segmentation results across four centers
图 3. 四中心典型分割可视化

4. 讨论

4.1. 核心发现与证据强度

本研究围绕“模型构建与多中心验证”建立了可量化证据链。最优配置(U-Net + ResNet50 + scSE + DiceBCE + Strong 增强 + 512 × 512)在同中心测试集取得 Dice 0.8760，在多中心验证集取得 Dice 0.8579；5-fold 与 LOCO 结果分别为 0.8741 ± 0.0019 和 0.8591 ± 0.0124，对应泛化差距 1.50%。这些指标共同指向一个结论：模型不仅在同域有效，在跨中心迁移时也保持相对稳定。

从效应量角度看，若以 MC Dice 作为泛化核心指标，Strong 增强较无增强提升 2.64 个百分点(0.8229 → 0.8493)，scSE 较 ResNet50 基线再提升 0.69 个百分点(0.8510 → 0.8579)；而负样本直接混入导致 3.85~5.87 个百分点下降。该结果说明，本研究的性能提升主要来自“正确策略叠加”，而非单一技巧或偶然最优轮次。

4.2. 数据策略反直觉结果的机制解释

本研究最具方法学价值的发现是：在样本极不平衡条件下，简单多中心混合训练并未形成稳定增益。数据构成上，Center 4 样本量明显高于其余中心(Center 1/2/3 分别约 372/429/575，Center 4 约 4271)，导致训练梯度长期由主中心分布主导。模型虽然“看到了”多中心数据，但未必真正学习到跨中心不变特征[8] [10] [11]。

该机制可由结果进一步印证：全中心混合训练(Exp-017)相比单中心训练(Exp-015)，MC Dice 仅轻微变化(0.8523 vs 0.8510)，但同中心 Test Dice 明显下降(0.8655 vs 0.8749)。这表明模型在跨域收益尚未形成前，先付出了同域性能代价。对于临床部署，若缺少分布对齐策略，简单堆叠数据可能并非最优路径[10] [11]。

负样本结果同样体现“任务特异性”。AFB 分割属于前景极稀疏任务，直接按 1:1、1:2、1:3 加入全零掩码样本后，MC Dice 分别降至 0.8009、0.7923 和 0.8125。其本质是优化目标被大量背景像素牵引，模型更容易学到“保守预测背景”的捷径，导致跨中心召回能力受损[5] [10]。

基于上述机制，后续数据策略应从“数量驱动”转向“分布驱动”：第一，采用按中心均衡采样或中心权重重标定，抑制主中心梯度垄断；第二，采用两阶段训练(主中心预训练 + 外中心适配微调)减少训练早期冲突；第三，在模型收敛后再引入困难负样本，而非在初始阶段大比例混入随机负样本。

4.3. 架构选择与参数效率的转化意义

结果显示，模型容量增加并不必然带来泛化提升。ResNet101 参数量更高(51.51 M)，但 Test Dice 和 MC Dice 分别仅为 0.8688 和 0.8387，低于参数更小的 EfficientNet-B3 (13.16 M；0.8751 和 0.8544)。这说明在中等规模医学数据上，过大容量更易放大过拟合和域特异噪声学习[5] [6]。

在注意力机制方面，scSE 在较低额外参数代价下实现 MC Dice 0.8579，为全流程最优。该结果支持“轻量增益优先”的部署思路：优先选择可解释、低开销、稳增益模块，而非复杂度高但收益不确定的结构[7]。

输入分辨率实验显示 512 × 512 是当前任务的实用平衡点。256 × 256 时细粒度形态信息受损(MC Dice 0.8435)；640 × 640 虽增加计算成本，但未获得相称精度提升。结合单图 0.18 s 推理时延，512 × 512 在性能与效率上更适合临床流程。

4.4. 泛化评估框架的临床解释价值

本研究同时报告同域测试、5-fold 稳定性和 LOCO 外域验证，目的在于避免“单一测试集高分”造成的部署乐观偏差。5-fold 反映训练可重复性，LOCO 反映“未见中心”条件下真实迁移能力，两者共同构成临床前验证的核心证据[10] [11]。

涂片级诊断符合率分析进一步完善了从像素到临床的证据链。在测试集阈值 ≥3 时，敏感性 96.30%、特异性 91.18%，表明分割模型在同域场景中具有接近人工判读的诊断一致性。在多中心验证集中，敏感性 94.44%、特异性 87.50%，虽略低于同域表现，但与像素级指标的跨中心衰减趋势一致(MC Dice 0.8579 vs Test Dice 0.8760)，提示样本级诊断效能的跨中心稳定性同样可控。

阈值分析揭示了筛查与确认两种临床场景下的策略选择。阈值 ≥ 1 时测试集敏感性达 100.00%，适用于“宁可错检、不可漏检”的初筛场景，但特异性降至 85.29%，意味着约 15%的阴性涂片可能被误报为阳性，需人工复核确认。阈值 ≥5 时特异性升至 97.06%，但敏感性降至 92.59%，可能漏检极低菌量的弱阳性样本。阈值 ≥ 3 在两项指标间取得较优平衡，可作为常规部署的默认推荐值，并可根据不同机构的临床优先级进行调整。

需指出的是，当前涂片级判定采用基于连通域计数的规则化策略，尚未引入模型置信度分布信息。后续可结合预测概率图构建加权评分机制，或采用基于涂片级标签的端到端训练策略，进一步优化样本级诊断效能。

泛化差距 1.50%可作为当前模型的风险沟通指标：该值并不代表“无风险部署”，而是提示模型在跨中心迁移时衰减幅度可控。对临床管理而言，这一指标有助于制定上限阈值、复核比例和复评周期。

4.5. 临床落地路径与风险控制

从像素级分割到临床样本级诊断的转化遵循逐级聚合逻辑：第一层为像素级分割，输出每个图像块的 AFB 掩码；第二层为视野级检出，通过连通域提取与面积过滤识别单个视野中的 AFB 候选目标；第三层为涂片级判定，依据全涂片检出目标数量与置信度分布，综合判定样本阳性或阴性。本研究 3.7 节的涂片级符合率分析初步验证了这一聚合路径的可行性，结果显示像素级分割精度可有效支撑样本级诊断判断。建议采用“AI 预标注–人工终审–持续复评”的闭环流程。AI 负责高通量候选区域标记与优先级提示，人工负责终审与疑难判读，系统按月回收分歧样本和低置信度样本回流用于再训练和模型迭代。该路径可在不替代临床决策责任的前提下提升效率与一致性[4] [8]。

为降低上线风险，建议设置三类监测指标：第一类为性能指标(Dice、Precision、Recall)；第二类为流程指标(平均复核时长、报告周转时间)；第三类为安全指标(高风险漏检率、中心间波动幅度)。当任一指标超过预设阈值时触发模型回滚或再训练。

此外，模型应用边界需在文中明确：本系统用于辅助筛查与分割提示，不替代最终诊断；对低质量制片、严重染色异常或非标准采集图像，应优先采用人工判读并记录为算法禁用场景[4] [12] [13]。

4.6. 局限性与后续改进方向

本研究为回顾性方法学研究，仍存在三方面局限。其一，外部中心规模仍可扩展，尚需纳入更多地区与设备条件验证稳健性。其二，当前以图像层面指标为主，尚未直接量化患者层面终点，如阳性检出率、人工复核负担和周转效率变化。其三，尚未完成前瞻性嵌入式试验，对真实临床行为改变的证据仍需补强。

后续工作建议聚焦两条主线：一是方法学上引入域泛化/域适应与不确定性建模，减少中心偏移影响；二是转化研究上开展前瞻性多中心实施研究，建立统一数据字典、统一复评规则和统一质量审计口径，形成“模型性能–流程收益–临床安全”一体化证据[8] [11]。

5. 结论

本研究完成了深度学习辅助痰涂片抗酸杆菌分割模型构建，并在多中心数据上进行了系统验证。最优配置 U-Net + ResNet50 + scSE + DiceBCE + Strong 增强 + 512 × 512 在测试集取得 Dice 0.8760，在多中心验证取得 Dice 0.8579。

5-fold 与 LOCO 验证结果分别为 0.8741 ± 0.0019 和 0.8591 ± 0.0124，泛化差距 1.50%，表明模型在跨中心部署场景具有较好稳健性。

涂片级诊断符合率分析表明，像素级分割输出经规则化聚合后，在测试集和多中心验证集中分别达到 93.44%和 90.48%的总符合率，初步验证了从分割模型到临床样本级诊断转化的可行性。研究同时提示，数据策略对性能影响显著：简单混合多中心数据与直接引入负样本并不一定带来收益。该结论对多中心医学 AI 建模具有普适参考价值。

在推理效率层面，模型可满足临床数字阅片的辅助时效需求，具有进一步开展前瞻性临床验证和流

程嵌入的现实基础。

作者贡献声明

徐小斐负责研究设计、算法开发、实验实施、数据分析和论文撰写；孙有湘负责多中心数据协调、图像标注质控和临床验证；王宏伟负责数据治理、统计分析和结果解读；李静负责图像预处理、标注审核和实验记录；张齐波负责外部中心数据收集与伦理协调；王清负责研究总体设计、方法论指导和论文审阅修订；高绪栋负责扫描系统技术支持与硬件环境配置；李培静负责图像采集流程优化与数据脱敏处理。所有作者均已阅读并同意最终稿件内容。

声　明

本研究方案经青岛大学附属医院医学伦理委员会审批通过(伦理批件号：QYFYWZLL30905)。

致　谢

感谢各合作医疗中心在病例管理、图像采集、数据标注与结果复核方面提供支持。感谢项目成员在算法实现、实验执行与数据治理方面的贡献。

参考文献

[1] 周林, 刘二勇, 孟庆琳, 等. 《WS 288—2017 肺结核诊断》标准实施后肺结核诊断质量评估分析[J]. 中国防痨杂志, 2020, 42(9): 910-915.

[2] 中国医学科学院病原生物学研究所, 中国疾病预防控制中心, 中国科学院地理科学与资源研究所. 全国结核分枝杆菌潜伏感染率估算专家共识[J]. 中国防痨杂志, 2022, 44(1): 4-8.

[3] 谢芳晖, 梁丽, 赵霞, 等. 肺结核患者痰标本采集的研究进展[J]. 中国防痨杂志, 2022, 44(9): 978-982.

[4] 中国防痨协会非结核分枝杆菌病分会. 非结核分枝杆菌病分子生物学诊断专家共识[J]. 中国防痨杂志, 2025, 47(8): 961-975.

[5] 曹玉红, 徐海, 刘荪傲, 等. 基于深度学习的医学影像分割研究综述[J]. 计算机应用, 2021, 41(8): 2273-2287.

[6] 李增辉, 王伟. 基于深度学习的医学图像分割方法研究进展[J]. 电子科技, 2024, 37(1): 72-80.

[7] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., *et al*., Eds., *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015, Springer, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[8] 张洪静, 褚洪迁, 孙照刚, 等. 我国结核病体外诊断科技成果创新转化现状分析[J]. 中国防痨杂志, 2024, 46(7): 743-749.

[9] 傅可言, 朱邦政, 叶健. 间质性肺疾病合并结核分枝杆菌感染的研究进展[J]. 中国防痨杂志, 2024, 46(7): 823-829.

[10] Del Carpio, C., Dianderas, E., Zimic, M., Sheen, P., Coronel, J., Lavarello, R., *et al*. (2019) An Algorithm for Detection of Tuberculosis Bacilli in Ziehl-Neelsen Sputum Smear Images. *International Journal of Electrical and Computer Engineering*, **9**, 2968-2981. https://doi.org/10.11591/ijece.v9i4.pp2968-2981

[11] Fu, H.T., Tu, H.Z., Lee, H.S., *et al*. (2022) Evaluation of an AI-Based TB AFB Smear Screening System for Laboratory Diagnosis on Routine Practice. *Sensors*, **22**, Article No. 8497. https://doi.org/10.3390/s22218497

[12] 夏辉, 赵雁林. 基于舌拭子的结核分枝杆菌复合群核酸检测方法在肺结核诊断和筛查中的应用前景与挑战[J]. 中国防痨杂志, 2025, 47(8): 976-980.

[13] 王苑柠, 杜宗敏. CRISPR/Cas 分子诊断技术在结核分枝杆菌耐药性检测中的研究进展[J]. 中国防痨杂志, 2025, 47(5): 666-672.