ICS 11.040.99
CCS C30

YY

中 华 人 民 共 和 国 医 药 行 业 标 准

YY/T XXXXX—202X

人工智能医疗器械 细胞病理图像辅助分析软件 算法性能测试方法

Artificial intelligence medical device—Computer assisted analysis software for cytopathologic images—Algorithm performance test methods

（征求意见稿）

本草案完成时间：2024 年 7 月 21 日

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家药品监督管理局　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由国家药品监督管理局提出。

本文件由人工智能医疗器械标准化技术归口单位归口。

本文件起草单位：

本文件主要起草人：

人工智能医疗器械 细胞病理图像辅助分析软件 算法性能测试方法

1 范围

本文件规定了采用人工智能技术的细胞病理图像辅助分析软件的算法性能测试方法。

本文件适用于采用人工智能技术对细胞病理图像进行后处理的辅助分析软件。

本文件不适用于细胞病理图像采集、前处理及过程优化类软件。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

YY/T 1833.1-2022 人工智能医疗器械质量要求和评价第1部分：术语

YY/T 1833.2-2022 人工智能医疗器械质量要求和评价第2部分：数据集通用要求

YY/T 1833.3-2022 人工智能医疗器械质量要求和评价第3部分：数据标注通用要求

YY/T 1858-2022 人工智能医疗器械 肺部影像辅助分析软件 算法性能测试方法

3 术语和定义

YY/T 1833.1、YY/T 1833.2、YY/T 1833.3界定的以及下列术语和定义适用于本文件。

3.1

数字病理图像 digital pathology lmages

通过数字传感技术与传统光学放大装置结合，在全自动显微镜或光学放大系统扫描采集得到的高分辨率的数字化图像，并能在模拟常规光学显微镜的计算机显示器上观察所扫描的病理玻片。

3.2

病理图像辅助分析 pathological images assisted analysis

利用人工智能技术对数字化的病理切片进行病灶区域或细胞结构分割与识别，提取图像纹理特征、图像颜色分布或细胞形态特征、空间分布特征，辅助开展病理类型分析、异常细胞检测、诊断等活动。

3.3

压力样本 stress sample

在某算法模型的标定范围内，特征容量极大或者极小的样本（如复杂细胞病变、异质性病变、稀有数据或引入玻片不包含的图像噪声等细胞病理图像），以确定算法模型的泛化性能、可靠性、稳定性。

[来源：YY/T 1858-2022，3.8，有修改]

3.4

压力测试 stress test

使用压力样本开展算法测试的过程。

[来源：YY/T 1858-2022，3.9]

4 测试要求

4.1 通则

细胞病理图像辅助分析软件的算法性能测试过程宜参照YY/T 1858-2022中4.1的要求，建立测试文档，给出明确规范的测试计划；如测试过程需要复测，应限定复测次数的上限（如不超过算法细胞分类或疾病诊断类型的数量），以避免算法对参考标准进行推测或针对性调优。

4.2 测试环境

算法性能的测试环境要求宜参照YY/T 1858-2022中4.2。

4.3 测试资源

4.3.1 测试前数据准备

在开始算法性能测试前，测试人员应先进行测试前数据准备。测试前数据准备流程图见图1。

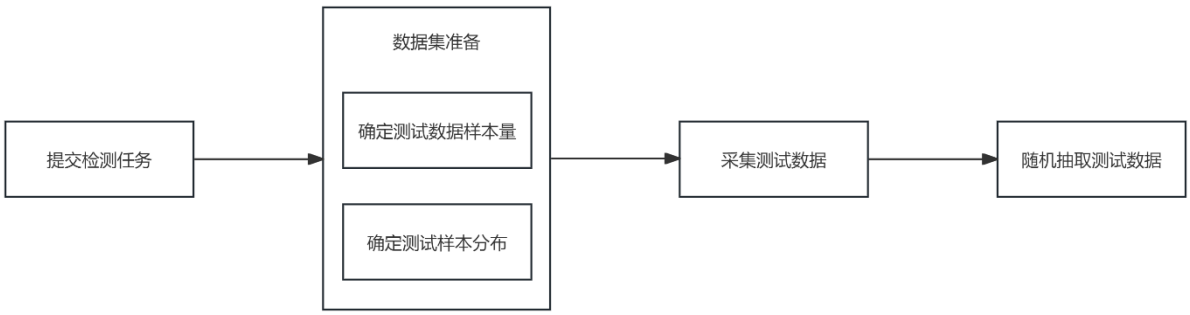


图1 测试前数据准备流程图

4.3.2 测试数据采集要求

4.3.2.1 图像数据采集

鉴于扫描设备、扫描参数、样本制作、人为操作等因素对细胞病理图像的数字采集过程产生影响，制造商应对采集设备、采集过程、人员培训等条件提出要求并建立图像数据采集操作规范。

注：附录A.1给出了图像数据采集的示例。

4.3.2.2 文本数据采集

为确保测试数据的完整性，宜采集和病理图像数据形成唯一映射的患者非敏感临床数据（如病理切片编号、年龄、诊断结论等）。

注：附录A.2给出了文本数据采集的示例。

4.3.2.3 数据脱敏和伦理

测试数据应获得相关伦理委员会批准或者豁免，源于临床脱敏数据。通过体检中心、社区筛查项目、科研项目等途径的数据收集行为同样应由伦理委员会进行审查和批准，保证数据脱敏、患者隐私安全和患者利益。

注：附录A.3给出了数据脱敏的方法示例。

4.3.2.4 测试数据样本量

细胞病理图像辅助分析软件的算法性能测试过程对样本量的最低要求宜参照YY/T 1858-2022中4.3.2的要求进行计算。

4.3.2.5 测试数据多样性

为保证测试数据具有充分的临床代表性，控制数据采集和收集过程导致的偏倚，数据应当尽可能覆盖到更多具有通用性的统计维度以提高算法模型效果的普适性。这些维度应包括：

a） 患者维度，考虑个体差异和地域差异的影响；

b） 场景维度，考虑不同的应用场景，如体检、筛查、门诊、手术和病理医学实验室等不同场所；

c） 设备与配置维度，考虑不同品牌、型号、不同分辨率及不同成像技术（如光学显微镜、电子显微镜）的病理成像设备。

d） 疾病构成维度，包括但不限于分型、分级、分期。

4.3.3 测试数据标注

本文件涉及的测试数据在标注时宜满足YY/T 1833.3-2022。数据集制造责任方应提供标注规则的来源，如世界卫生组织（WHO）、国内外公开发表的各类疾病的专家共识或诊断标准，确保标注标签准确、完整。

注：附录A.4给出了测试数据标注的示例。

4.3.4 扩增数据

细胞病理图像辅助分析软件的算法性能测试过程中，如有必要产生扩增数据，应参照执行YY/T 1858-2022中4.3.4的要求，并且经人工确认后对扩增数据进行交付使用。

4.4 测试平台

如使用测试平台进行算法测试，测试平台宜满足YY/T 1858-2022中4.4的要求。

4.5 测试指标与通过原则

测试人员应根据产品的预期用途和使用场景，以及产品的技术特性和风险分析，在测试文档中明确列出客观和可定量的测试指标。制造商应给出产品相对应各指标的标称值及允差或上下限。

对于产品应用场景测试指标，测试人员应根据产品预期用途和使用场景，确定适用的测试指标：

——若最终应用目的主要是给出详细细胞分类的软件产品，则应用场景测试指标宜选择适用于评价细胞类型识别的指标；

——若最终应用目的主要是给出具体疾病诊断类型的软件产品，则应用场景测试指标宜选择适用于评价病理诊断准确性的指标。

对于产品算法质量特性指标，测试人员应根据产品的技术特性和风险分析，确定适用的整体评估指标，作为算法质量特性的判定依据。

细胞病理图像辅助分析软件的算法性能测试通过准则，应包括产品应用场景测试指标和算法质量特性指标两个方面均通过。

4.6 测试流程

测试人员宜参照YY/T 1858-2022中4.6的要求，根据测试文档开展测试活动，并形成完整的测试记录。

4.7 测试结果

测试人员宜参照YY/T 1858-2022中4.7的要求，形成测试结果的描述文档。

5 算法性能测试方法

5.1 算法应用场景的测试方法

5.1.1 细胞病理图像分割场景

5.1.1.1 概述

细胞病理图像分割包括区域分割和细胞结构分割两个维度。区域分割是通过人工智能技术训练获取识别细胞病理图像中正常细胞和疑似病变细胞区域或不同形态的细胞区域，具有自动区域分割的产品输出的是感兴趣细胞区域分割图(Segmentation of region of interest tissue, SROI)，SROI可用于可疑病变细胞的定位、特定细胞形态和数量占比测量计算和细胞病变程度分析。细胞结构分割（Segmentation of cellular structure, Seg）是指细胞核、细胞膜和细胞质等结构的分割，细胞结构分割结果通常用于细胞数量、轮廓大小和形态分布等量化指标计算。

5.1.1.2 测试步骤

细胞病理图像分割场景测试的参考标准为满足本文件4.3.2测试数据采集要求和4.3.3测试数据标注要求的所有内容后输出的分割标签，测试人员在测试计划中说明算法结果和参考标准结果的匹配方式和匹配阈值，匹配方式和匹配阈值的界定由产品制造商声称并提供。

在细胞病理图像分割场景下，算法性能测试按如下步骤进行：

a) 向待测算法输入测试集，输出算法结果，算法结果的格式宜与参考标准兼容；
b) 算法分割的目标区域与参考标准分割的目标区域的 SROI 或 Seg 的性能指标，按 5.1.1.3 和 5.1.1.4 描述的公式进行评价指标计算；

注1：以病灶区域或目标对象为单元的，整个集合的计算结果取平均值作为最终结果。
注2：以病例为单元的，计算每个病例结果，然后对病例集合的计算结果取平均值作为最终结果。

5.1.1.3 客观一致性评价指标

客观一致性评价指标目标区域采用重合指标来评估分割模型，包括Dice系数（Dice coefficient，Dice）和Jaccard一致性系数（Jaccard similarity coefficient）。

5.1.1.3.1 Dice 系数

表示算法预测出的目标区域和参考标准区域相交面积的两倍占两者目标区域总和的比值，用Dice表示，表达式见公式(1)：

$$Dice = \frac{2\times|S_{pr}\cap S_{gt}|}{|S_{pr}|+|S_{gt}|} \quad (1)$$

式中：

S_{pr}——算法分割的目标区域；

S_{gt}——参考标准分割的目标区域。

5.1.1.3.2 Jaccard 一致性系数

表示算法预测出的目标区域和参考标准区域之间的相似系数，用Jaccard表示，表达式见公式(2)：

$$Jaccard = \left|\frac{PS\cap GT}{PS\cup GT}\right| \quad (2)$$

式中：

PS——算法预测分割结果；

GT——参考标准分割。

5.1.1.4 客观差异性评价指标

客观差异性评价指标采用95%豪斯多夫距离（95% Hausdorff Distance，HD95）、平均对称表面距离（Average Symmetry Surface Distance，ASSD），都是基于表面距离的评估指标，用于度量算法预测结果和参考标准分割的最大差异。

5.1.1.4.1 95%豪斯多夫距离

反映分割结果中存在的离群区域，用HD95表示，表达式见公式(3)：

$$HD95(X,Y) = percentile_{95\%}\left\{\left\{\min_{y\in Y} d(x,y)\,|x\in X\right\},\left\{\min_{x\in X} d(y,x)\,|y\in Y\right\}\right\} \quad (3)$$

式中：

$d(x,y)$——表示自动分割区域的边界点x和参考标准分割区域边界点y之间的欧氏距离；

$\left\{\min_{y\in Y} d(x,y)\,|x\in X\right\}$——表示对于自动分割区域中的每个边界点，计算其到参考标准区域边界点的最短欧氏距离所构成的集合；

$\left\{\min_{x\in X} d(y,x)\,|y\in Y\right\}$——表示对于参考标准区域中的每个边界点，计算其到自动分割区域边界点的最短欧氏距离所构成的集合；

$percentile_{95\%}$——表示将上述所有计算的距离值进行排序，对自动分割区域到参考标准区域的所有最短距离，取95%分位点；对参考标准区域到自动分割区域的所有最短距离，同样取95%分位点。

5.1.1.4.2 平均对称表面距离

用来衡量自动分割区域边界与参考标准区域边界之间的平均距离，用ASSD表示，表达式见公式(4)：

$$ASSD(X,Y)=mean\left\{\left\{\min_{y\in Y} d(x,y)\,|x\in X\right\},\left\{\min_{x\in X} d(y,x)\,|y\in Y\right\}\right\} \quad (4)$$

式中：

$d(x,y)$——表示自动分割区域的边界点x和参考标准分割区域边界点y之间的欧氏距离；

$mean$——表示求均值。

5.1.2 细胞类型识别场景

5.1.2.1 概述

细胞类型的识别，包括观察目标细胞的数量及形态的变化，或对某些类型细胞群的量变或质变的准确识别，从而辅助医生进行疾病诊断。

细胞类型识别场景测试的参考标准为满足本文件4.3.2测试数据采集要求和4.3.3测试数据标注要求的所有内容后输出的细胞标签，测试人员在测试计划中说明算法输出的细胞分类范围及参考标准的细胞分类范围，以及每一类细胞与参考标准相比达到的匹配阈值，细胞分类范围及每一种细胞匹配阈值的界定由产品制造商声称并提供。

5.1.2.2 测试步骤

在细胞类型识别场景下，算法性能测试按如下步骤进行：

a) 细胞类型识别场景的算法输出结果，是以每个分割的细胞图像为维度输出具体的细胞类型判断结果，因此本场景的测试，要求必须是在通过细胞病理图像分割场景的测试后进行；

b) 向待测算法输入测试集，输出算法结果，算法结果的格式宜与参考标准兼容；

c) 比较算法预测细胞类型与参考标准分类，输出细胞类型识别的真阳性、假阳性、真阴性、假阴性结果，构建混淆矩阵。二分类混淆矩阵见表 1。

表1 二分类混淆矩阵

分类		算法模型	
		阳性	阴性
参考标准分类	阳性	真阳性 （true positive, TP）	假阴性 （false negative, FN）
	阴性	假阳性 （false positive, FP）	真阴性 （true negative, TN）

d) 本文件涉及的细胞类型识别实际多为多分类问题，而多分类实际可转化为二分类问题，参考标准分类为 i 类与其他非 i 类别的混淆矩阵简化，见 YY/T 1833.1-2022 表 A.1 和表 A.3；

e) 按照 5.1.2.3-5.1.2.5 描述的公式进行评价指标计算。

注：5.1.2.3-5.1.2.5描述的公式中，TP、FP、TN、FN代表的含义是从单个细胞的维度与参考标准比较得出的结果。

5.1.2.3 精确度

精确度用Pre表示，表达式见公式(5)：

$$Pre=\frac{TP}{TP+FP}*100\% \quad (5)$$

5.1.2.4 召回率

召回率用Rec表示，表达式见公式(6)：

$$Rec=\frac{TP}{TP+FN}*100\% \quad (6)$$

5.1.2.5 F1 度量

F_1度量表达式见公式(7)：

$$F_1=2*\left(\frac{\mathrm{Pre}*\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}}\right) \quad (7)$$

式中：

Pre——表示细胞类型识别的精确度；

Rec——表示细胞类型识别的召回率。

5.2 算法质量特性与测试方法

5.2.1 泛化能力

制造商应根据产品预期用途和部署环境，对产品研发使用的训练集与真实世界全新样本之间的差异进行分析，形成文档，作为配置测试集的依据。实际测试中，宜参照附录A要求的数据采集和标注方法建立多样性与变化性的独立测试集，对算法的泛化能力进行验证。

5.2.2 鲁棒性

制造商应根据产品风险分析和临床部署环境特征，利用数据扰动、生成对抗网络等技术产生对抗样本，并采用对抗样本开展的算法性能测试，分析各指标的变化情况，形成鲁棒性研究资料。可采取的测试方法包含但不限于：

——对抗样本测试：生成或使用对抗样本，即那些经过特意设计以欺骗算法的输入数据，来测试算法的鲁棒性。观察算法在面对这些恶意修改的数据时的表现；

——噪声样本注入：在输入数据中注入染色剂分布不均匀、气泡、灰尘、制片污染、细胞碎片等噪声样本，测试算法对不同类型和强度噪声的处理能力；

——异常样本测试：在数据中引入训练数据以外的细胞类型或亚细胞类型的异常样本，评估算法处理这些数据时的稳定性和性能表现。

5.2.2.1 面向硬件变化的对抗测试方法

测试人员应考虑病理成像硬件设备兼容性、参数设置、实验参数的多样性等，收集或模拟生成更多的数字病理图像数据，作为对测试集的扩充以满足多样性要求，验证算法面对病理成像采集硬件设备的鲁棒性。参数设置应考虑：物理分辨率、曝光时间、放大倍率、染色剂、染色设备等。模拟生成的图像数据不应影响标注结论。面向硬件变化的对抗测试方法应包括：

a) 跨成像设备测试：在不同品牌和型号的病理成像设备上进行测试，包括不同分辨率、不同成像技术（如光学显微镜、电子显微镜）的设备，以评估算法对于从不同硬件来源的图像数据的处理能力；
b) 成像参数变化测试：改变成像过程中的关键参数设置，如物理分辨率、曝光时间、放大倍率等，来评估算法对于图像质量变化的适应性。这种测试帮助确定算法在实际应用中遇到参数设定不一致时的鲁棒性；
c) 实验参数变化测试：由于染色剂和染色设备的差异可能对图像产生显著影响，应通过使用不同的染色剂和染色设备对样本进行处理，并评估算法处理这些图像的能力，从而测试算法在面对实验参数变化时的稳定性；
d) 跨硬件平台测试：测试算法在不同品牌、不同规格的硬件上的表现，包括 CPU、GPU、TPU 等不同类型的处理器；
e) 硬件资源限制下的性能：验证算法在不同的资源限制条件下（如内存大小、存储容量、处理速度）的性能表现；
f) 跨设备兼容性：确保算法能够在来自不同制造商和不同技术规格的设备上运行，而不会出现兼容性问题；
g) 模拟环境测试：使用虚拟机或容器技术在同一硬件上模拟不同硬件环境，测试算法在这些环境下的性能和适应性。

5.2.2.2 面向数据预处理（相机图像处理和软件前处理）的对抗测试方法

测试人员应通过考虑和应用各种数据预处理步骤（如背景去除、裁剪、增强等）来收集或模拟生成多样化的图像数据，以此扩充测试集并验证算法的鲁棒性。这些步骤旨在模拟现实世界中可能遇到的数据预处理情况，同时确保模拟数据的标注结果保持一致，从而全面评估算法对前处理变化的适应性。主要步骤应包括：

a) 应用多种预处理操作：在原始图像数据集上施加包括背景去除、图像裁剪、图像增强、彩色/灰度模式转换等多种软件前处理步骤，以反映现实世界的多样性；

b) 模拟生成挑战性图像样本：使用图像处理工具或编程脚本来模拟各种前处理效果，生成具有对抗性的测试图像，同时确保这些图像的标注与原始数据保持一致；
c) 评估算法性能：通过 5.1.1.3.1、5.1.2.4、5.1.2.5 关键性能指标来评估算法在经过预处理的图像数据集上的表现，与算法在原始未处理图像数据集上的性能进行对比分析；

5.2.2.3 面向欺骗攻击的对抗测试方法

测试人员可使用白盒攻击验证模型是否能抵御恶意欺骗攻击。主要步骤应包括：
a) 白盒攻击：在白盒攻击场景下，攻击者对模型的结构和参数有完全的了解，使用投影梯度下降（projected gradient descent, PGD）算法生成最大范数有限的扰动，这种扰动设计得足够微小（例如不超过 8/256），以至于对人类观察者而言几乎不可察觉；
b) 插入扰动到原始图像：将通过 PGD 等算法生成的扰动加入原始图像中，创建一组修改后的图像数据集；
c) 模型测试：使用含扰动的图像数据集对目标模型进行测试，观察模型在面对这些经过精心设计的欺骗性扰动的图像时的表现，测试标准包括模型对这些图像的分类准确率、误报率等性能指标。

5.2.2.4 压力测试

5.2.2.4.1 压力样本的选取

测试人员宜从测试集中选取压力样本，并开展压力测试，压力样本不应影响医生判断。
细胞病理图像辅助分析算法的压力测试样本选取，考虑的特征应包括：
a) 异质性病变样本；
b) 复杂病变的混合型样本；
c) 稀有数据样本；
d) 复杂的病理学表现样本；
e) 引入图像噪声的样本。

5.2.2.4.2 压力测试的步骤

压力测试方法应包括以下步骤：
a) 性能评估：使用选定的压力样本对算法进行测试，关注算法在处理这些极端或复杂条件下的表现，特别是其准确性、稳定性和处理速度等关键性能指标；
b) 结果分析：对测试结果进行深入分析，识别算法在处理特定类型的压力样本时可能遇到的问题和挑战，如识别精度下降、处理时间增加等；

5.2.3 重复性

重复性测试方法应包括以下步骤：
a) 选择或生成测试数据：挑选重复或生成相同的目标数据分布和疾病类型的样本数据作为测试集；
b) 多次执行算法：对选定的相同样本数据，采用同一版本的算法进行多次测试，测试次数不宜低于 3 次，记录每次的输出结果；
c) 结果比较分析：对算法多次运行的输出结果进行比较，评估一致性。

5.2.4 一致性

5.2.4.1 确定参考标准

使用多位具有资深经验的病理医生组成的专家组对细胞切片进行常规显微镜判读，形成参考标准。确保专家组中至少包含3位病理医生，并明确判定决策机制。

5.2.4.2 样本量要求

选择的测试数据集样本量根据产品预期的特异度和敏感度来确定，用灵敏度计算阳性组的样本量，用特异度计算阴性组的样本量，阳性组/阴性组的最大值是单次测试样本量的最低要求。

阳性组/阴性组样本量的估算公式为：

$$n = \frac{Z_{1-\alpha/2}^{2} P(1-P)}{\Delta^{2}} \quad (8)$$

式中：

n——阳性组/阴性组样本量；

$Z_{1-\alpha/2}$——为标准正态分布的分位数；

P——为灵敏度或特异度的预期值；

Δ——为 P 的允许误差大小，一般取 P 的 95%置信区间宽度的一半，常用的取值为 0.05—0.10。

5.2.4.3 测试方法选择

一致性测试方法选择应参考以下内容：

a) 对于预期用于分割的模型：采用 5.2.1 描述的方法，通过比较算法输出结果与参考标准标记之间的客观一致性和客观差异性来衡量一致性；
b) 对于预期用于分类的模型：采用 5.2.2 描述的方法，建立混淆矩阵并计算 Kappa 系数，来评估算法分类结果与参考标准之间的一致性。
c) 对于预期用于分割和分类多任务的模型：采用 5.2.1 和 5.2.2 描述的方法，每个任务分别评估算法结果与参考标准之间的一致性结果，若涉及相同测试指标需进行平均性能计算。

5.2.5 分析效率

5.2.5.1 定义测试起止点

分析效率测试的起止点，应满足以下内容：

a) 起点：数据开始导入算法的时刻；
b) 终点：对于大多数应用，终点为算法导出全部结果的时刻。对于辅助分析类产品，终点特定为生成算法通知的时刻。

5.2.5.2 明确临床典型病例要素

明确临床典型病例时考虑的要素应包括以下内容：

a) 规定参与测试的图像数量，如 100 张切片；
b) 明确细胞病理切片的制片方式，如涂片或抹片的方式；
c) 指定成像方式，如使用光学显微镜；
d) 确定染色方式，如巴氏染色、瑞氏染色。

5.2.5.3 执行测试

分析效率的测试方法应包括以下步骤：

a) 根据 5.2.5.2 描述的明确临床典型病例的要素，准备测试数据集；
b) 测量从起点到终点的总时间，确保测试环境稳定，以便结果具有可比性；
c) 分析处理时间，考虑其对临床流程的影响；
d) 与现有技术或手工处理时间进行比较，以评估算法带来的效率改进。

5.2.6 错误分析

5.2.6.1 错误分析的方法内容

错误分析的方法应包括以下内容：

a) 分割场景的错误分析：对于图像分割任务，根据感兴趣区域（ROI）的尺寸，细致分析分割结果的准确性。这包括评估算法在不同尺寸的 ROI 上的表现，以及识别可能导致分割错误的因素；

b) 多细胞分类场景的错误分析：在细胞分类任务中，对每一种分类结果进行详细分析，特别是假阴性和假阳性结果。这有助于识别算法在特定类型上可能存在的弱点；
c) 针对个体病例的性能评估：对每个病例进行算法性能指标的计算，从而评估算法是否存在对特定病例的偏倚。这可以揭示算法在处理不同类型病例时的泛化能力和局限性；
d) 在对抗测试和压力测试中的应用：在进行对抗测试和压力测试时，同样应采用上述方法进行错误分析。这包括对分割精度、分类结果的假阴性和假阳性，以及算法对个体病例性能的评估，从而全面理解算法在极端条件下的表现和潜在的弱点。

5.2.6.2 错误分析的测试步骤

错误分析的测试方法应包括以下步骤：

a) 数据准备：根据测试需要，准备相应的数据集，包括分割、分类任务的标准数据集，以及对抗测试和压力测试所需的特定数据；
b) 执行测试：运行算法，收集错误数据（如分割不准确的区域、分类的假阴性和假阳性结果等）；
c) 数据分析：利用统计和分析方法，对收集到的错误数据进行详细分析，识别错误模式和原因；
d) 报告编写：将分析结果整理成详细的测试报告，包括错误类型、频率、潜在原因，以及算法对不同病例的性能表现和偏倚分析。

附 录 A
（资料性）
测试数据采集和标注示例

A.1 图像数据采集

A.1.1 采集设备

图像采集设备参数的参考示例见表A.1。

表A.1 图像采集设备参数的参考示例

性能	指标	
光源	明场光源，具备高色彩还原度。	
XY载物台	带自动控制功能的XY电动载物台	
物镜转换	物镜数量≥2，建议配备10×或20×、40×物镜。	
物镜要求（N.A.值）	10×	N.A.≥0.4
	20×	N.A.≥0.8
	40×	N.A.≥0.9
	100×	N.A.≥1.4
扫描模式	单层对焦/智能对焦，能识别和跳过无样本的区域	
图片像素	一般不低于100万像素，具有数字放大功能；显示器上可缩放图像。	
注：N.A.值：数值孔径是判断物镜性能的重要因素，它与分辨率成正比。		

A.1.2 采集参数

扫描范围：满足诊断医生进行诊断及鉴别诊断为原则，保证扫描范围满足诊断需要。
扫描分辨率：满足病理医生诊断需求。
存储格式：满足任意人工智能算法的通用要求。

A.1.3 采集人员

经过规范化培训能够保证满足图像质量要求的数字病理图像扫描人员。

A.1.4 图像数据质量审核

图像质量审核依据数据集制造责任方引用的相关临床诊疗指南或专家共识；图像质量审核一般包含主观评价和客观评价两个方面。

图像质量的主观评价由中级及以上职称的医生执行。

图像质量的客观评价包含但不限于如下内容：

（1）细胞病理切片成像扫描范围的完整性；

（2）图像亮度、对比度和色彩空间；

（3）图像清晰度。

数据集制造责任方对审核过程中发现的错误予以纠正，对无法纠正的错误数据采取退回或去除操作。

A.2 文本数据采集

制定文本数据的统一采集规则，采集内容包括但不限于如下内容：

（1）患者基本信息：姓名、病案号、性别、年龄、所在医院、检查日期、样本编号。

（2）电子病历项目：现病史（如治疗方式、服用药物）、既往史（如有无明确诊断）。

A.3 数据脱敏和伦理隐私

A.3.1 内容

扫描人员使用高清扫片仪采集图像时、工作人员在录入电子病历时会一并采集到病人信息，在样本进入数据集时，被脱敏信息包括：患者基本信息（姓名、ID）。患者年龄、性别、检查日期、所在医院、病史不需要脱敏。保留患者样本编号仅用作区分患者使用。

A.3.2 方法

使用合法稳定的工具读取关联文件内所携带的病人信息，参照拟遵循的要求，判定敏感字段信息，确定后利用程序自动将需要脱敏的信息进行清洗；或采用人工删除敏感信息的方式进行脱敏处理。

A.3.3 可追溯性

为满足数据后期与其他临床资料对照及后续跟进研究的需要，保证病例可追溯，故在数据脱敏之前对所提取的患者信息进行加密备份，用于后续追溯使用。

A.4 测试数据标注要求

A.4.1 总则

本文件涉及的测试数据标注过程执行YY/T 1833.3-2022的规定，标注规则参照世界卫生组织（WHO）当前对各个系统肿瘤的分类及诊断标准，以及国内外核心期刊公开发表的各类疾病的专家共识和标准，对数据集每一个样本赋予准确的完整标签。包括分割标签、细胞标签和疾病标签。

分割标签：包括所有目标对象所在的分割区域，且不包含任何其他无关区域，如其他无关区域会影响到目标对象的分割区域标注时，需同时输出无关区域的分割标签。

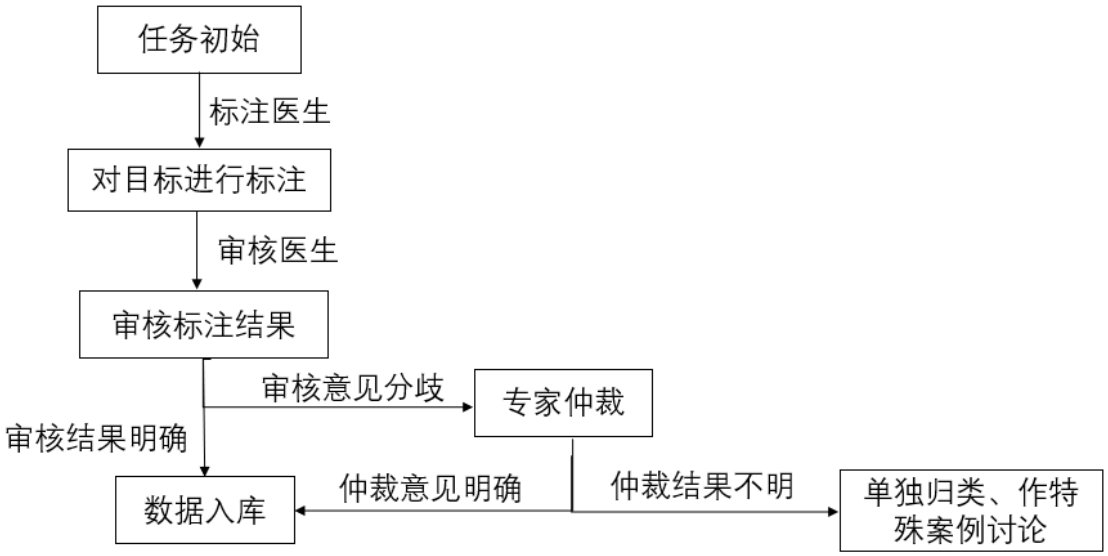
细胞标签：以细胞为目标的项目将细胞切割完整，包含完整的细胞核，标注精确的细胞边界。

疾病标签：指符合审核标准的病理医生做出的具体疾病诊断。

A.4.2 标注操作规范

整体的标注和质控机制采用“两标一审一仲裁”模式，即一张数字病理图像由两名病理医生进行数据标注，综合标注意见，由审核医生审核决定是否采用。考虑不同环节的工作量和人员资质的差异，为提高标注的准确性，降低假阴性率，标注医生一般为主治以上病理医生，审核医生一般为具有专业权威性的高年资主治或以上级别的医生。若标注医生和审核医生意见偏移，则移交更高级别仲裁专家予以裁定。数据标注流程示意图见图A.1。

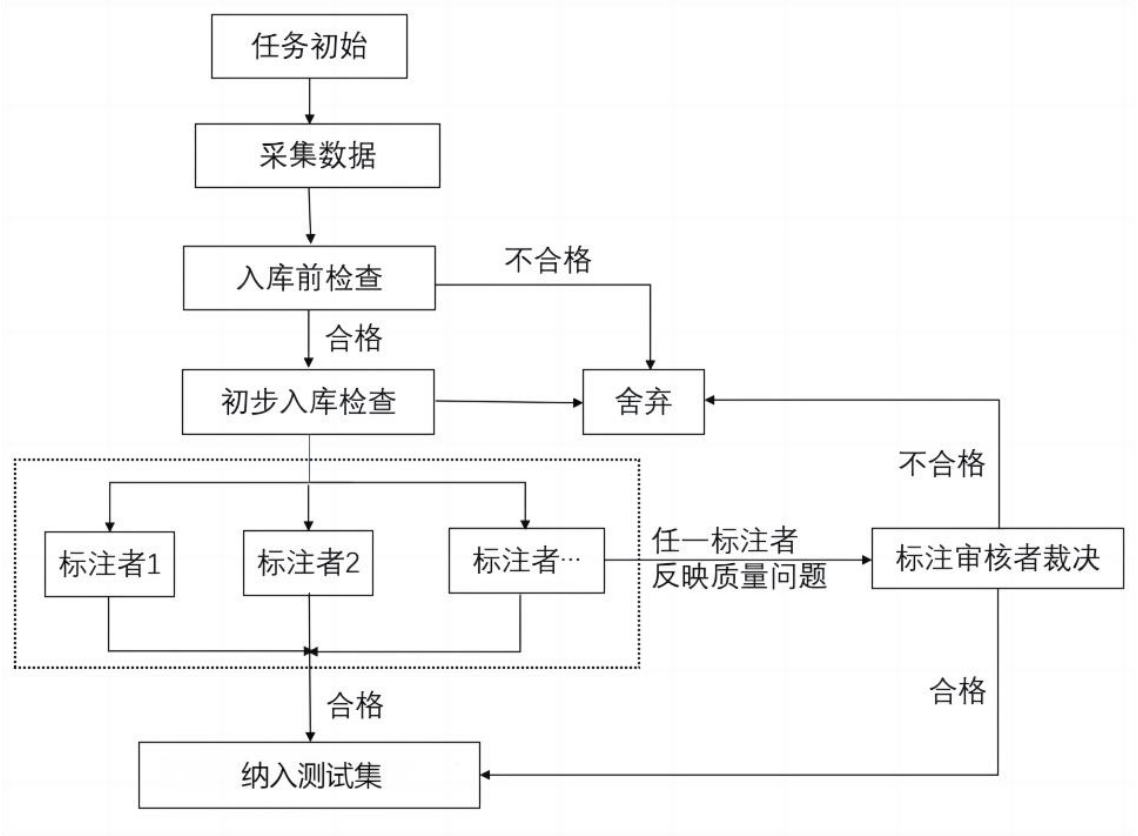
图A.1 数据标注流程示意图



A.4.3 数据入库原则

数据采集处理完成后经过一系列质控环节，包括入库前检查、初步入库检查、标注过程检查，方可纳入测试集（详见图A.2）。入库前检查主要对数据相关指标进行固定规则的简易筛查（如图像分辨率等）；初步入库检查是由研究者对入库数据进行的进一步检查（如特定形态结构位置是否合乎要求）；标注过程检查是由标注者对可能影响标注的其他数据质量问题进行报告，最终由标注审核者进行裁决。数据入库流程示意图见图A.2。

图A.2 数据入库流程示意图



参 考 文 献

[1] 国家药品监督管理局医疗器械技术审评中心.深度学习辅助决策医疗器械软件审评要点.[Z].2019.

[2] 中华人民共和国中央人民政府.新一代人工智能发展规划[Z].2017-07-20.

[3] 中华人民共和国原国家卫生和计划生育委员会.人工智能辅助诊断技术管理规范[Z].2017-02-17.

[4] 中华人民共和国原国家卫生和计划生育委员会.人工智能辅助诊断技术临床应用质量控制规范[Z]. 2017-02-17

[5] 国家药监局医疗器械技术审评中心.人工智能医疗器械注册审查指导原则[Z], 2022年第8号.

[6] 国家药品监督管理局医疗器械技术审评中心.病理图像人工智能分析软件性能评价审评要点.[Z].2023

[7] 刘恩彬,蔺亚妮,王慧君,李承文，汝昆． 血液肿瘤的综合诊断[J]． 中华血液学杂志,2016,37(1)：83-86.

[8] 《宫颈液基细胞学人工智能辅助诊断数据集标注规范与质量控制专家共识(2022版)》编写组.宫颈液基细胞学人工智能辅助诊断数据集标注规范与质量控制专家共识（2022版）[J]. 中华病理学杂志,2022,51(12)：1205-1209.

[9] 中国病理医师协会数字病理与人工智能病理学组,中华医学会病理学分会数字病理与人工智能工作委员会,中华医学会病理学分会细胞病理学组. 宫颈液基细胞学的数字病理图像采集与图像质量控制中国专家共识[J]. 中华病理学杂志,2021,50(04)：319-322.

[10] 白求恩精神研究会检验医学分会,中华医学会检验医学分会血液体液学组,中国医学装备协会检验医学分会基础检验设备学组. 人工智能辅助外周血细胞形态学检查的中国专家共识[J]. 中华检验医学杂志,2023,46(03)：243-258.

[11] 中国医师协会检验医师分会造血与淋巴组织肿瘤检验医学专家委员会. 造血与淋巴组织肿瘤检验诊断报告模式专家共识[J]. 中华医学杂志,2016,96(12)：918-929.

[12] Palmer L, Briggs C, McFadden S, et al. ICSH recommendations for the standardization of nomenclature and grading of peripheral blood cell morphological features[J]. Int J Lab Hematol. 2015 Jun;37(3):287-303. doi: 10.1111/ijlh.12327. Epub 2015 Mar 2. PMID: 25728865.

[13] Kratz A, Lee SH, Zini G, et al; International Council for Standardization in Haematology. Digital morphology analyzers in hematology: ICSH review and recommendations[J]. Int J Lab Hematol. 2019 Aug;41(4):437-447. doi: 10.1111/ijlh.13042. Epub 2019 May 2. PMID: 31046197.

[14] Aeffner F, Zarella MD, Buchbinder N, et al. Introduction to Digital Image Analysis in Whole-slide Imaging: A White Paper from the Digital Pathology Association[J]. J Pathol Inform. 2019 Mar 8;10:9. doi: 10.4103/jpi.jpi_82_18. Erratum in: J Pathol Inform. 2019 Apr 24;10:15. doi: 10.4103/2153-3539.259372. PMID: 30984469; PMCID: PMC6437786.

[15] Abels E, Pantanowitz L, Aeffner F, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. 2019 Nov;249(3):286-294. doi: 10.1002/path.5331. Epub 2019 Sep 3. PMID: 31355445; PMCID: PMC6852275.

[16] James B, Mary G, Christian W, et al. TNM classification of malignant tumors[M]. John Wiley & Sons, 2017.

[17] Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence[J]. Lancet Oncol. 2019 May;20(5):e253-e261. doi: 10.1016/S1470-2045(19)30154-8. PMID: 31044723; PMCID: PMC8711251.

[18] Colling R, Pitman H, Oien K, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice[J]. J Pathol. 2019 Oct;249(2):143-150. doi: 10.1002/path.5310. Epub 2019 Jul 18. PMID: 31144302.

[19] Center for Devices and Radiological Health in U.S. Department of Health and Human Services Food and Drug Administration, Technical Performance Assessment of Digital Pathology Whole Slide Imaging Device[Z],2015-4-20.