ICS 点击此处添加 ICS 号
CCS 点击此处添加 CCS 号

中华人民共和国国家标准

GB/T XXXXX—XXXX

人工智能医疗器械 超声影像辅助分析软件 算法性能测试方法

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

立项草案稿

XXXX－XX－XX 发布 XXXX－XX－XX 实施

国家市场监督管理总局
国家标准化管理委员会 发布

人工智能医疗器械　超声影像辅助分析软件　算法性能测试方法

1　范围

本文件描述了采用人工智能技术的超声影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甲状腺与乳腺超声B型成像进行实时处理或后处理的辅助分析软件产品。

本文件不适用于超声影像前处理软件。

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

YY/T 1907—2023　人工智能医疗器械　冠状动脉CT影像处理软件　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界定的术语和定义适用于本文件。

3.1

语义分割　semantic segmentation

将图像中每个像素标注语义类别标签的过程。

3.2

实例分割　instance segmentation

在语义分割的基础上，对同一语义类别的不同对象实例进行区分。

3.3

聚类分析　cluster analysis

一种无监督学习技术，主要用于将样本划分成多个类别或簇，使得在同一类别或簇内的样本相似性较高，不同类别或簇之间的样本相似性较低。

注：聚类分析的目标是在不需要预先定义类别的情况下，发现样本中的内在结构和模式。

3.4

重识别　re-identification；ReID

利用计算机视觉技术识别并匹配不同图像或视频序列中特定目标的技术，其主要目的是在图像或视频序列中对感兴趣目标进行检索。

3.5

超声造影　contrast-enhanced ultrasound；CEUS

在常规超声检查的基础上，通过外周静脉注射超声造影剂，实时动态观察组织的微血管灌注信息，以提高病变的检出率及良恶性的鉴别。

4　算法性能测试要求

4.1　概述

超声影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；如测试过程需要复测，应对复测次数进行限制，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的要求。在进行重复性和再现性测试时，需要注意使用差异化的主要硬件环境，如不同的满足算法运行最低要求的CPU和GPU型号。如果算法训练和测试是在不同人工智能算法框架下进行的，则测试内容应包括不同框架环境下的再现性测试。

4.3 测试资源

4.3.1 测试集要求

测试集要求宜参照YY/T 1907—2023中4.3.1的要求。

4.3.2 数据采集要求

4.3.2.1 影像质量要求

超声影像质量要求如下：

——使用超声设备采集的原始图像，应去除个人敏感信息，不能改变图像中有效超声区域内的内容；回顾性数据需满足有效超声区域内无测量标记、文字等干扰信息；

——明确使用的超声影像类型或组合图像类型组合，包括灰阶图像、多普勒图像、弹性图像、造影图像等；组合图像应针对同一病灶、组织或器官进行分析；

——病灶图像至少包括 2 张有特征的不同方向切面，宜留存最长径切面和与之垂直的切面，切面应尽量反应病灶超声特征，如不规则形状、边缘模糊、边界不光整、钙化等；

——对于超声检查没有异常的组织或器官，也要留存超声图像以表明没有缺失或表明已对患者做过全面的超声检查。

4.3.2.2 测试集样本量要求

对于分割、检测和分类场景下的算法性能测试，测试集样本量宜参照YY/T 1858—2022中4.3.2计算；测量的一致性分析所需要的测试集样本量可参考YY/T 1907—2023中附录B；病灶关联的聚类和重识别技术以及病灶跟踪没有关于最小样本量普遍接受的经验法则，制造商宜根据产品需求明确最小样本量确定依据。

4.3.2.3 测试集多样性要求

测试数据集宜体现产品适用范围和临床使用场景内的样本多样性，包括但不限于成像设备维度、成像参数维度、扫查方式维度、患者维度。

成像设备维度考虑以下内容：

——超声设备品牌及其具体型号；

——超声设备探头型号。

成像参数维度考虑以下内容：

——频率；

——增益；

——深度。

扫查方式维度考虑以下内容：

——扫查部位；

——扫查角度，如横切扫查、纵切扫查、其它任意角度扫查。

患者维度考虑以下内容：

——性别、年龄、地域、疾病进展期等因素的多样性；

——产品适用范围，考虑病变类型，如下：

- 病灶位置、数量、大小、深度、距离；
- 分级；
- 病理亚型。

4.3.2.4 扩增数据要求

扩增数据要求宜参照YY/T 1907—2023中4.3.2.4的要求。其中白盒扩增可以对超声图像进行网格形变、弹性形变和添加散斑噪声等。

4.3.3 测试工具要求

采用仿组织超声体模的测试方法，仅适用于软件功能验证，不适用于对整个软件的泛化能力、鲁棒性等进行完整评价。如适用，算法测试使用的体模、标准器宜参照YY/T 1858—2022中4.3.5的要求。

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 测试指标与通过准则

测试指标与通过准则的选取宜参照执行YY/T 1858—2022中4.5的一般要求；测试指标应根据产品的适用功能选择，宜按照切面、病灶、患者等不同的统计维度进行计算。

测试计划应根据产品预期用途制定，描述病灶的定义、测量方式、分级/分类方式、病例总体结论的确立规则等适用的信息，作为开展测试的依据。

4.6 测试流程

测试流程宜参照YY/T 1858—2022中4.6的一般要求。

4.7 测试结果

测试结果描述宜参照YY/T 1858—2022中4.7的一般要求。

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 分割

5.1.1.1 基于像素的评价方法

5.1.1.1.1 像素准确率

像素准确率（Pixel Accuracy，PA）是指在所有测试集图像中，被正确分类的像素数量除以像素总数，计算方法见公式（1）：

$$PA = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K}\sum_{j=0}^{K} p_{ij}} \quad \cdots\cdots (1)$$

式中：

PA——像素准确率；

K ——前景类别数量；

p_{ij}——参考类别是i被预测为类别j的像素数量。

5.1.1.1.2 平均像素准确率

平均像素准确率（Mean Pixel Accuracy，MPA）需要首先计算每一类被正确分类的像素数除以该类别的像素总数，然后对所有类别求平均，计算方法见公式（2）：

$$MPA = \frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K} p_{ij}} \quad \cdots\cdots (2)$$

式中：

MPA——平均像素准确率。

5.1.1.2 基于区域的评价方法

基于区域的评价方法是一种全局度量方式，与基于像素的评价方法相比，其对局部差异不敏感。宜采用Jaccard系数或Dice系数计算分割区域与参考标准区域的重叠程度，具体计算过程参照YY/T 1858—2022中5.1.2.4的方法。Jaccard系数与Dice系数呈正相关，用公式（3）表示：

$$\mathrm{Dice}=\frac{2\times\mathrm{Jaccard}}{\mathrm{Jaccard}+1} \quad \cdots\cdots (3)$$

注：Jaccard系数又称为交并比（Intersection over Union，IoU）。

在对每张图像中具体的每个标记进行评估时，宜参照YY/T 1858—2022中5.1.1的方法，在确定Jaccard系数或Dice系数的匹配阈值后，可以计算召回率、精确度、F1度量、平均精确度、平均精确度均值和构造自由响应受试者操作特征曲线。由于语义分割算法无法输出分割区域的类别概率值，因此没有算法阈值可以调整，不能使用平均精确度、平均精确度均值和自由响应受试者操作特征曲线评价算法性能，而实例分割算法可以采用平均精确度、平均精确度均值和构造自由响应受试者操作特征曲线进行评价。

与平均精确度指标偏重于预测区域的精确性不同，平均召回率指标更偏重于预测的全面性，其定义为：不同Jaccard系数或检测框数量限制下召回率的平均值。平均召回率更能反应小目标的检出效果。

在对整张超声图像进行分割评价时，宜采用全图分割质量（Panoptic Quality，PQ）。全图分割质量定义为“识别质量”（Recognition Quality，RQ）与“分割质量”(Segmentation Quality，SQ)的乘积，计算方法见公式（4）：

$$\mathrm{PQ}=\mathrm{RQ}\times\mathrm{SQ}=\frac{2\times\mathrm{TP}_{t=0.5}}{2\times\mathrm{TP}_{t=0.5}+\mathrm{FP}_{t=0.5}+\mathrm{FN}_{t=0.5}}\times\frac{\sum_{(p,g)\in\mathrm{TP}_{t=0.5}}\mathrm{Jaccard}(p,g)}{|\mathrm{TP}_{t=0.5}|} \quad \cdots\cdots (4)$$

式中：

PQ ——全图分割质量；

t ——匹配阈值，使用Jaccard系数关于该阈值计算TP，FP，FN；

RQ ——识别质量；

SQ ——分割质量；

TP ——真阳性区域的数量；

FN ——假阴性区域的数量；

FP ——假阳性区域的数量；

(p,g)——匹配上的预测区域和参考区域。

在超声扫查视频中对每个病灶的分割结果进行评价，可以将视频的每一帧作为独立图像进行计算，如考虑病灶在视频中的连续性，则将匹配方法扩展到视频上计算三维连通域和Jaccard系数，然后计算相关的评价指标。

5.1.1.3 基于边界的评价方法

基于边界的评价方法也是一种全局度量方式，通过分割区域的边界评价感兴趣区域的结构或形状的相似或不相似性，可使用双向豪斯多夫距离（Hausdorff Distance，HD），其计算方法参照YY/T 1858—2022中5.1.2.6，但其对异常值比较敏感，因此还可以使用95%双向豪斯多夫距离（95% Hausdorff Distance，HD95）和平均双向豪斯多夫距离（Average Hausdorff Distance，AHD）。

5.1.1.3.1 95%双向豪斯多夫距离

95%双向豪斯多夫距离表示预测区域的边界点集和参考区域的边界点集的最小距离的最大值的第95百分位数，计算方法见公式（5）：

$$\mathrm{HD95}(X,Y)=\max_{k95\%}\{\max_{x\in X}\min_{y\in Y}d(x,y),\max_{y\in Y}\min_{x\in X}d(x,y)\} \quad \cdots\cdots (5)$$

式中：

$\mathrm{HD95}(X,Y)$——95%双向豪斯多夫距离；

X ——预测区域的边界点集；

Y ——参考区域的边界点集；

$d(x,y)$ ——X和Y两个点集任意两点之间的距离；

$k95\%$ ——第95百分位数。

5.1.1.3.2 平均双向豪斯多夫距离

在计算任一单向豪斯多夫距离时，使用计算每个点最小距离的平均值代替最大值，计算方法见公式（6）：

$$\mathrm{AHD}(X,Y)=\max\{\frac{1}{N}\sum_{x\in X}\min_{y\in Y}d(x,y),\frac{1}{M}\sum_{y\in Y}\min_{x\in X}d(x,y)\} \quad (6)$$

式中：

$\mathrm{AHD}(X,Y)$——平均双向豪斯多夫距离；

N ——预测区域的边界点集X的点的数量；

M ——参考区域的边界点集Y的点的数量。

5.1.1.4 交互式分割的评价方法

利用交互式分割算法可以分割出算法完全遗漏的区域或调整已分割区域的边界，以进行下一步分类、测量等任务，帮助形成完整的病例检查报告。交互方式可以是点击、划线、画框等。

在交互式分割的自动测试过程中，通过比较参考标注和预测区域的差异，下次交互将在最大误差区域的中心进行，采用达到目标Jaccard系数所需要的平均交互次数作为评价指标，目标Jaccard系数宜设置为85%或90%，表示为NoC@85、NoC@90，其值越小表示交互分割的效率越高。

5.1.2 检测

检测与基于区域的分割评价方法基本相同，其主要区别是匹配计算方法不同，以Jaccard系数为例，检测的Jaccard系数使用矩形框计算，而分割的Jaccard系数使用逐像素计算。检测的评价方法宜参照YY/T 1858—2022中5.1.1的方法，在确定匹配阈值后，计算召回率、精确度、F1度量、平均精确度和平均精确度均值，也可以画出自由响应受试者操作特征曲线。

5.1.3 分类

分类的评价方法参照YY/T 1858—2022中5.1.3的方法：在二分类任务（包括多分类转换为二分类）中，根据确定的分类阈值构造混淆矩阵，计算灵敏度、特异度、漏检率、阳性预测值、阴性预测值、准确率、约登指数、Kappa系数；如果分类阈值未确定，可以通过计算受试者操作特征曲线（Receiver Operating Characteristic Curve，ROC Curve）的曲线下面积（Area Under the Curve，AUC）值评价算法整体性能；在每类同等重要的多分类任务中，一般选取概率值最大的标签作为预测类别，构造多分类混淆矩阵，可以计算准确率和Kappa系数。

在ROC曲线有相交时，可以比较ROC曲线下选定区域的面积，即计算局部AUC值。局部AUC值可以通过约束灵敏度、特异度的范围来计算，也可以应用其他客观约束条件获得。

在二分类时，当正负样本数量很不平衡时，可以使用G-mean指数，G-mean指数是正样本的召回率和负样本的召回率的综合指标，计算方法见公式（7）：

$$\mathrm{G-mean}=\sqrt{\frac{\mathrm{TP}}{\mathrm{TP+FN}}\times\frac{\mathrm{TN}}{\mathrm{TN+FP}}}=\sqrt{\mathrm{Sen}\times\mathrm{Spe}} \quad (7)$$

式中：

TP ——真阳性样本的个数；

FN ——假阴性样本的个数；

TN ——真阴性样本的个数；

FP ——假阳性样本的个数；

Sen——灵敏度；

Spe——特异度。

马修斯相关系数（Matthews Correlation Coefficient，MCC）也是一种用于评估不平衡数据集分类性能的指标，MCC的计算方法见公式（8）：

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP+FP})\times(\mathrm{TP+FN})\times(\mathrm{TN+FP})\times(\mathrm{TN+FN})}} \quad (8)$$

MCC的取值范围在-1到+1之间，其中：+1表示完美预测，0表示随机预测，-1表示预测与实际观察完全不一致。

在一些超声实时扫查或检查视频分析任务中，需要对视频中的每一帧进行分类，可以计算帧级准确率、帧级召回率、帧级精确度和帧级F1度量，也可以以视频为单位，计算视频级的这些指标，计算方法

和常规分类任务计算方法一样。考虑到视频中每一帧的类别在时间维度上有关联，还可以通过比较帧到帧的时间预测标签曲线和标注标签曲线的一致性进行评价，如计算豪斯多夫距离。

5.1.4 测量

组织或病灶的大小采用测量径线长度表示，径线测量的评估包括径线定位的准确性和径线测量的一致性。径线定位可以使用图像处理或机器学习中的分割、检测方法，首先确定径线的两个端点坐标，然后计算预测点坐标与参考点坐标的欧式距离，在确定匹配阈值后，只有两个端点坐标都预测正确，则径线定位正确，可以计算每条径线的召回率、精确度、F1度量。在考虑待测量目标大小以及参考标注的主观差异因素时，径线定位也可以使用径线端点相似度（Object Keypoint Similarity，OKS）描述，计算方法见公式（9）：

$$OKS = \frac{1}{N}\sum_{i=1}^{N} \exp(-\frac{d_i^2}{2s^2k_i^2}) \quad \cdots\cdots (9)$$

式中：

OKS——径线端点相似度；

N ——所有径线端点数量；

d ——端点预测坐标与参考坐标的欧式距离；

s ——待测目标的尺度因子，等于待测区域的面积或体积的平方根；

k ——端点的归一化因子，取参考标注的标准差，值越大表示越难标注。

径线测量的评价方法可以参照YY/T 1858—2022中5.1.2.8的尺寸测量方法，计算相对误差绝对值的平均值，也可以参照YY/T 1907—2023中5.1.4.1的一致性分析，选用Bland-Altman分析、组内相关系数或Pearson相关系数计算方法。

在用径线表示体积时，需要所有径线都定位正确，并且每条径线的测量值与参考值的误差在规定范围内，则表示体积测量正确，统计测试集中测量正确的比例，表示体积测量的准确率。

5.1.5 切面关联

在超声影像检查中，同一个病灶会有多张不同的切面影像，而同一张切面影像上可能会存在多个病灶，在病灶大小测量、随访等应用场景中，需要先采用聚类或重识别技术将病灶切面进行关联。

5.1.5.1 聚类

当需要确定病例的病灶数量以及每个病灶的所有切面时，可以采用聚类方法对病灶切面进行分组关联，聚类方法可选用以下指标和计算方法。

5.1.5.1.1 兰德指数

兰德指数（Rand Index，RI）的计算方法见公式（10）：

$$RI = \frac{a+b}{C_{n_{sample}}^2} = \frac{2(a+b)}{n_{sample}(n_{sample}-1)} \quad \cdots\cdots (20)$$

式中：

$C_{n_{sample}}^2$ ——所有样本可能配对的总数；

n_{sample} ——所有样本的总数；

A ——在真实标签和预测聚类中都归为同一簇的样本对的数量；

B ——在真实标签和预测聚类中归为不同簇的样本对的数量。

RI取值范围为[0,1]，值越大表示聚类结果与真实情况越相似。然而，兰德指数并不能保证随机标签分配将获得接近于零的值，特别是如果簇的数量与样本数量处于相同的数量级。

5.1.5.1.2 调整兰德指数

调整兰德指数（Adjusted Rand Index，ARI）解决了在聚类结果随机产生的情况下兰德指数不接近于0的问题，计算方法见公式（11）：

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad \cdots\cdots (31)$$

式中：

ARI ——调整兰德指数；

E[RI] ——兰德指数的期望值；

max(RI)——兰德指数的最大值。

5.1.5.2 重识别

如果只对某一切面的病灶感兴趣，则不需要对所有病灶切面进行关联分组，可以采用重识别方法对感兴趣切面进行关联，重识别方法可选用累计匹配特性曲线和平均精确度均值进行评价。

5.1.5.2.1 累计匹配特性曲线

已知所有查询样本，为了计算累计匹配特性曲线，首先把查询结果按相似程度从大到小排序，则前k个命中的准确性top_k定义如下：

$$\mathrm{top\_k}=\begin{cases}1, & \text{前k个结果中有与查询样本相同标签的样本}\\ 0, & \text{否则}\end{cases}$$

可以看出，top_k是平移阶跃函数。累计匹配特性曲线的横坐标表示k，k取值可以不连续，如k等于1、2、5等，纵坐标表示在所有查询样本中top_k等于1的样本数量，与查询样本总数的比值，即准确率。在实际应用时可以根据场景要求确定k值。

5.1.5.2.2 平均精确度均值

在重识别场景下的平均精确度与其他场景（如检测）的平均精确度不一样，由于模型提供了每个样本与查询样本相似性由大到小的排序，通过在排序列表中的每个位置计算精确度和召回率，可以得到精确度-召回率曲线。重识别的平均精确度计算方法见公式（12）：

$$\mathrm{AP}=\frac{\sum_{k=1}^{M}\mathrm{p(k)}\times\mathrm{gt(k)}}{\mathrm{N_{gt}}} \quad (42)$$

$$\mathrm{gt(k)}=\begin{cases}1, & \text{如果排序列表中第k个样本与查询样本标签相同}\\ 0, & \text{否则}\end{cases}$$

式中：

AP ——平均精确度；

M ——模型返回的排序列表的长度；

$\mathrm{N_{gt}}$ ——排序列表中与查询样本有相同标签的样本数量；

p(k)——在排序列表位置k的精确度，其值等于前k个样本中与查询样本标签相同的样本数量除以k。

计算测试集中所有查询样本的平均精确度的平均值，就是重识别的平均精确度均值。

5.1.6 病灶跟踪

在超声实时扫查中，病灶跟踪需要实时检测或分割出所有病灶位置并给出每个病灶的唯一标识，常用的评价指标有：多病灶跟踪准确率、多病灶跟踪精确度、身份标识F1分数和高阶跟踪准确率。

5.1.6.1.1 多病灶跟踪准确率

多病灶跟踪准确率（Multi-Lesion Tracking Accuracy，MLTA）是一个结合了假阳性、假阴性和身份标识切换次数的综合评价指标，计算方法见公式（13）：

$$\mathrm{MLTA}=1-\frac{\sum_t \mathrm{FN_t+FP_t+IDSW_t}}{\sum_t \mathrm{GT_t}} \quad (53)$$

式中：

MLTA——多病灶跟踪准确率；

$\mathrm{FN_t}$ ——第t帧分割或检测的假阴性数量；

$\mathrm{FP_t}$ ——第t帧分割或检测的假阳性数量；

$\mathrm{IDSW_t}$——第t帧分割或检测的目标身份标识发生切换的次数；

$\mathrm{GT_t}$ ——第t帧真实的病灶数量。

MLTA的取值范围是$(-\infty,1]$，值越大表示跟踪效果越好。

5.1.6.1.2 多病灶跟踪精确度

多病灶跟踪精确度（Multi-Lesion Tracking Precision，MLTP）描述的是所有正确匹配的目标和其对应真实目标之间的平均差异，计算方法见公式（14）：

$$\mathrm{MLTP}=\frac{\sum_{t,i}d_{t,i}}{\sum_{t}c_{t}} \quad \cdots\cdots (64)$$

式中：

MLTP——多病灶跟踪精确度；

$d_{t,i}$ ——第t帧中，第i个预测目标与其匹配的真实目标之间的匹配程度或距离；

c_t ——第t帧中，正确匹配的目标数量。

由此可见，MLTP只是分割或检测精度的衡量标准，其值与$d_{t,i}$的计算方式有关，如采用Jaccard系数表示匹配程度，则MLTP值越大表示病灶跟踪的位置越精确。

5.1.6.1.3 身份标识 F1 分数

身份标识F1分数评估预测的轨迹与真实轨迹之间身份标识的一致性，计算预测轨迹和真实轨迹之间身份标识匹配的精确度和召回率的F1分数。计算方法见公式（15）、（16）、（17）：

$$\mathrm{IDF1}=\frac{2\times\mathrm{ID_{Precision}}\times\mathrm{ID_{Recall}}}{\mathrm{ID_{Precision}}+\mathrm{ID_{Recall}}} \quad \cdots\cdots (75)$$

$$\mathrm{ID_{Recall}}=\frac{\mathrm{IDTP}}{\mathrm{IDTP+IDFN}} \quad \cdots\cdots (16)$$

$$\mathrm{ID_{Precision}}=\frac{\mathrm{IDTP}}{\mathrm{IDTP+IDFP}} \quad \cdots\cdots (17)$$

式中：

IDF1 ——身份标识F1分数；

$\mathrm{ID_{Recall}}$ ——身份标识召回率；

$\mathrm{ID_{Precision}}$ ——身份标识精确度；

IDTP ——正确匹配的真实轨迹的数量；

IDFP ——没有匹配上的预测轨迹的数量；

IDFN ——没有匹配上的真实轨迹的数量。

轨迹的匹配方式可参考本文件中的5.1.1.3。

5.1.6.1.4 高阶跟踪准确率

高阶跟踪准确率（Higher Order Tracking Accuracy，HOTA）是结合且平衡定位精度和关联精度的比较全面的评价指标。高阶跟踪准确率的设计旨在：

a） 为跟踪算法评估提供单一评分，该评分能公平地结合所有不同的跟踪评估方面；

b） 评估长期高阶跟踪关联；

c） 可以分解为多个子指标，从而能够分析跟踪算法性能的不同组成部分。

计算方法见公式（18）、（19）、（20）：

$$\mathrm{HOTA}=\int_0^1\mathrm{HOTA_\alpha}d\alpha\approx\frac{1}{19}\sum_{\alpha\in\{0.05,0.1,\ldots0.9,0.95\}}\mathrm{HOTA_\alpha} \quad \cdots\cdots (18)$$

$$\mathrm{HOTA_\alpha}=\sqrt{\frac{\sum_{c\in\{TP\}}A(c)}{|TP|+|FN|+|FP|}} \quad \cdots\cdots (89)$$

$$\mathrm{A(c)}=\frac{|\mathrm{TPA(c)}|}{|\mathrm{TPA(c)}|+|\mathrm{FNA(c)}|+|\mathrm{FPA(c)}|} \quad \cdots\cdots (90)$$

式中：

HOTA ——高阶跟踪准确率；

HOTA_α——在检测或分割的匹配阈值是α时的高阶跟踪准确率；

TP ——检测或分割的真阳性病灶数量；

FN ——检测或分割的假阴性病灶数量；

FP ——检测或分割的假阳性病灶数量；

A(c) ——被正确检测或分割出的病灶c的关联分数；

TPA(c)——被正确检测或分割出的病灶c的真阳性关联病灶数量；
FNA(c)——被正确检测或分割出的病灶c的假阴性关联病灶数量；
FPA(c)——被正确检测或分割出的病灶c的假阳性关联病灶数量。

5.1.7 随访评估

参照YY/T 1858—2022中5.1.4所述，输入病例不同时间节点的数据，使用切面关联方法进行病灶匹配，根据各时间节点建立病灶量化指标的变化曲线。预测曲线与参考曲线的一致性可以使用分割应用场景中基于边界的评价方法，如计算双向豪斯多夫距离。

5.1.8 多功能组合

参照YY/T 1858—2022中5.1.4，测试人员应对具有多种功能的产品进行分步评价，如：

a) 首先对病灶进行分割，计算分割的评价指标；
b) 其次对分割正确的病灶进行分类或切面关联，计算分类、聚类分析或重识别的评价指标；
c) 最后对关联正确的病灶测量大小，计算测量误差等。

5.2 算法质量特性与测试方法

5.2.1 泛化能力

宜参照YY/T 1858—2022中5.2.1，结合本文件4.3.2.3的要求和制造商规定的产品适用范围，对测试样本进行抽取和组合，观测算法在测试集的不同子集上的性能差异，子集一般体现设备及具体型号差异、成像参数差异、扫查方式差异以及患者差异。

5.2.2 鲁棒性

宜参照YY/T 1858—2022中5.2.2，基于制造商规定的产品适用范围进行扩展测试，分析各指标的变化情况，形成鲁棒性研究资料。

对抗测试宜考虑超声设备差异和数据采集条件差异导致的图像变化，可使用模拟仿真产生的数据，在人工确认后用于测试。

测试人员宜选取压力样本开展压力测试，压力样本不应影响医生判断，压力样本的选取可考虑、但不限于以下特征：

a) 斑点噪声较多；
b) 边界不清或不规则；
c) 伪影，如声影、后方回声增强、混响等。

5.2.3 重复性

宜参照YY/T 1858—2022中5.2.3，对同一版本的算法至少进行3次重复测试，观察测试结果是否变化。

5.2.4 再现性

如果算法训练和部署使用不同的深度学习框架，需要测试算法在不同框架下的再现性。如产品不涉及指定硬件，则应在不同的满足算法最低运行环境的设备上测试，观察测试结果的再现性。

5.2.5 一致性

如算法输出的中间结论具有参考标准，宜参照YY/T 1858—2022中5.2.4，使用5.1的适用方法对中间结论进行验证，对算法输出的中间结论和产品输出的最终结论进行比较。

示例：如当产品判断病例的病灶类型时，算法判断的该病灶各个切面的结果是否与该病灶结论一致。

5.2.6 效率

宜参照YY/T 1858—2022中5.2.5，以测试数据开始导入的时刻作为起点，以算法输出结果的时刻作为终点。临床典型病例需约定图像格式、尺寸和数量，如病灶分割和分类是独立的算法，还需要约定病灶数量。如果病灶分割和分类是独立任务且不连续执行（如需要对分割结果进行交互修改），应分别记

录完成病灶分割和分类的效率。如涉及交互分割算法，宜分别记录算法第一次给出分割结果的时间和交互的时间（从第一次给出分割结果到最终确认分割的时间），每次交互中间不停留。

病灶跟踪的处理模式分为实时和离线两种：

a） 实时跟踪模式：对视频进行逐帧处理，当前帧只能利用过去帧信息；

b） 离线跟踪模式：可以利用当前帧的前后多帧信息对当前帧进行处理。

所以离线跟踪模式不能用于实时超声扫查。病灶跟踪的效率主要用每秒处理帧数（Frames Per Second, FPS）来衡量，其直接反映算法运行速度。FPS指标在记录测试环境的同时，还要记录纯模型推理速度和含前、后处理的模型推理速度。

在超声检查的实时应用场景中，应保证算法效率不能影响正常的诊疗流程。

5.2.7 算法错误统计

宜参照YY/T 1858—2022中5.2.6，对算法测试错误结果进行统计分析，如：

——不同层级（切面、病灶、病例）的假阴性、假阳性；

——在分割、检测和分类应用场景下，测试人员宜根据单切面的病灶数量，对结果进行分组统计；

——在分割、检测和病灶跟踪应用场景下，测试人员宜根据匹配阈值，对结果进行分组统计；

——在切面关联和病灶跟踪应用场景下，测试人员宜根据单病例的病灶数量，对结果进行分组统计；

——在分类应用场景下，测试人员宜根据单病灶的切面数量，对结果进行分组统计。

附 录 A
（资料性）
特定条款的指南和方法说明

A.1 通用指南

对于采用人工智能技术的超声影像处理软件的算法性能质量评价，测试要求贯穿整个测试过程，第4章对测试要求进行了详细阐述，第5章对算法性能测试方法进行了描述，按照应用场景测试和质量特性两个角度展开。

A.2 特定章条的指南和方法说明

以下内容与正文的章条相对应，提供推荐方法或举例。

对于5.1.3超声检查视频的每一帧进行分类的应用场景，如在肝脏超声造影检查中，对每一帧进行造影时相分类，造影时相和持续时间的关系：

——动脉期：通常指超声造影剂注射后的 10 s～20 s 到 30 s～45 s；

——门脉期：通常指造影剂注射后的 30 s～45 s 持续至 2 min；

——延迟期：通常指从门脉期结束持续到超声造影剂微泡从血循环中基本清除，约 4 min～6 min。

对于5.1.7病灶跟踪应用场景中，匹配方法与单张图像检测或分割应用场景的匹配方法不同，病灶跟踪在匹配时一般以最大化评价指标为目标，如在高阶跟踪准确率HOTA的匹配计算过程中，使用匈牙利算法来选择匹配集合，其首要目标是最大化真阳性病灶的数量，其次是最大化真阳性病灶集合的关联得分均值，第三是最大化真阳性病灶集合的定位相似性均值。每个预测目标和真实目标的潜在匹配分数的计算方法见公式（A.1）：

$$\mathrm{MS}(i,j) = \{\frac{1}{\epsilon} + \mathrm{A_{max}}(i,j) + \epsilon \mathrm{S}(i,j) \ \text{ if } \mathrm{S}(i,j) \geq \alpha\} \quad \cdots\cdots (A.1)$$

式中：

$\mathrm{MS}(i,j)$——真实目标i和预测目标j的潜在匹配分数；

ϵ ——平衡评分公式中不同组成部分的权重；

α ——定位相似度的匹配阈值；

$\mathrm{A_{max}}$ ——在定位不是双射匹配时的最大关联分数；

$\mathrm{S}(i,j)$ ——真实目标i和预测目标j的定位相似度，如Jaccard系数。

由上式可知，高阶跟踪准确率HOTA的匹配方法不仅与定位精度有关，还与病灶身份标识关联程度有关，通过在每一帧执行最优匹配过程，考虑了所有可能的轨迹关联，确保HOTA能够准确地反映跟踪算法的整体性能。

对于5.1.8随访评估应尽量采用相同的检查方法、体位和仪器设置，对于需要监测大小和形态变化（如边缘是否变得不规则、内部回声是否发生改变等）的病变，应进行准确、可重复的测量，最好采用相同的测量方法和参考点。对于接受手术等治疗的患者，应首先考虑算法的适用范围，再进行随访评估判断治疗效果。

附 录 B
（资料性）
分割和检测算法平均精确度的计算方法

B.1 精确度-召回率曲线的平均精确度计算方法

针对分割、检测算法，参照YY/T 1858—2022中5.1.1，确定标记匹配方法和匹配阈值，可以计算召回率、精确度和F1度量，改变算法阈值设置，生成精确度-召回率曲线。根据精确度—召回率曲线，常用的有两种计算平均精确度的方法：11点插值法和所有点插值法。11点插值法计算简单，但结果没有所有点插值法精确。

11点插值法：对一组11个召回率下的精确度进行平均，11个召回率的值是召回率从0到1以0.1为间隔均匀采样获得的，其对应精确度的值是通过取召回率大于当前召回率的最大精确度值插值获得的。在一些场景下，也可以选取更多的点插值，如召回率以0.01为间隔均匀采样，则有101个点计算平均精确度。

所有点插值法：通过对所有点插值，平均精确度可以表示为精确度-召回率曲线下的面积，可以减少曲线波动的影响。精确度-召回率曲线下的面积可以采用近似方法计算，一般有两种方式：矩形规则和梯形规则，两种方法计算的结果略有差异。

B.2 不同匹配阈值的平均精确度

分割和检测算法的固定匹配阈值通常选取Jaccard系数等于0.5，对于Jaccard系数大于0.5的预测其重要程度是一样的，因此仅采用单个匹配阈值预测会在评估指标中引入偏差。解决这一问题的有效方法是使用一系列匹配阈值，计算每个匹配阈值下的平均精确度，最后取所有匹配阈值下的平均精确度的平均值作为最终的平均精确度。以使用Jaccard系数匹配方法为例，匹配阈值从0.5开始，以0.05为间隔均匀选取，直到0.95结束，共10个匹配阈值，计算10个匹配阈值下的平均精确度，表示为AP@[0.50:0.05:0.95]。常用的AP50和AP75指标，表示匹配阈值为0.5和0.75时的平均精确度，匹配阈值越高表明对算法输出的精确程度要求越高。

参 考 文 献

[1] GB 10152—2009 B型超声诊断设备

[2] GB/T 25000.12—2017 系统与软件工程 系统与软件质量要求和评价（SQuaRE） 第12部分：数据质量模型

[3] YY/T 0767—2009 超声彩色血流成像系统

[4] YY/T 1833.4—2023 人工智能医疗器械 质量要求和评价 第4部分：可追溯性

[5] YY/T 1833.5—2024 人工智能医疗器械 质量要求和评价 第5部分：预训练模型

[6] 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（国家药品监督管理局 2019年第7号）.

[7] 国家药品监督管理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（2022年第8号）.

[8] 国家药品监督管理局医疗器械技术审评中心.医疗器械软件注册审查指导原则（2022年第9号）.

[9] 国家药品监督管理局医疗器械技术审评中心.肺结节CT图像辅助检测软件注册审查指导原则（国家药品监督管理局 2022年第21号）.

[10] 国家药品监督管理局医疗器械技术审评中心.影像超声人工智能软件（流程优化类功能）技术审评要点（国家药品监督管理局 2023年第23号）.

[11] 浙江大学, 中国食品药品检定研究院,海军军医大学第二附属医院.人工智能医疗器械性能评价通用方法专家共识（2023）[J]. 协和医学杂志,2023,14(3):494-503. DOI:10.12290/xhyxzz.2023-0137.

[12] Minaee S, Boykov Y, Porikli F, et al. Image segmentation using deep learning: A survey[J]. IEEE transactions on pattern analysis and machine intelligence,2021,44(7):3523-3542.

[13] Luiten J, Osep A, Dendorfer P, et al. Hota: A higher order metric for evaluating multi-object tracking[J]. International journal of computer vision, 2021, 129：548-578.

[14] 陈敏华,严昆,戴莹,等.肝超声造影应用指南（中国）（2012年修改版）[J].中华超声影像学杂志,2013,22(8)：696-722. DOI:10.3760/cma.j.issn.1004-4477.2013.08.018.

[15] 刘睿峰,夏宇,姜玉新.人工智能在超声医学领域中的应用[J].协和医学杂志, 2018, 9(5)：453-457. DOI：10.3969/j.issn.1674-9081.2018.05.015.

[16] 王权,王浩,张超,等.超声诊断类人工智能医疗器械测试方法研究[J].中国医疗设备, 2023, 38(4)：35-39 https://doi.org/10.3969/j.issn.1674-1633.2023.04.007.

[17] 李胜利,秦越,谭光华,等.医学超声人工智能的应用与挑战[J].中华医学超声杂志（电子版）,2023, 20(01)：1.