ICS 35.240
CCS L67

团　　体　　标　　准

T/EGAG 016—2022

智慧医疗影像辅助诊断平台质量评估规范

Specification for quality evaluation of intelligent medical image aided diagnosis platform

2022－12－14 发布　　2022－12－14 实施

广东省电子政务协会　　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专利的责任。

本文件由广东省电子政务协会归口。

本文件起草单位：工业和信息化部电子第五研究所、中山大学中山眼科中心、中山大学附属第一医院。

本文件主要起草人：高岩、许振豪、尤俊、徐欣、潘聪、林浩添、林铎儒、项毅帆、赖伟翊、尚元君、匡铭、彭穗、王伟、肖晗。

智慧医疗影像辅助诊断平台质量评估规范

1 范围

本标准规定了智慧医疗影像辅助诊断平台的质量评估体系和评估方法。

本标准适用于指导智慧医疗影像辅助诊断平台开发方、用户方以及第三方等相关组织对平台的质量开展评估工作。

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。

GB/T 25000.51-2016 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第51部分：就绪可用软件产品（RUSP）的质量要求和测试细则

ISO 14155:2020 Clinical investigation of medical devices for human subjects — Good clinical practice

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能 artificial intelligence

人工智能被广泛定义为制造智能机器的科学和工程，尤其是智能计算机程序。人工智能可以使用不同的技术，包括数据统计分析、专家系统、机器学习和深度学习等，通常体现于被认为具有人类智能的系统功能中，如推理和学习。

3.2

智慧医疗影像辅助诊断平台 Intelligent medical image aided diagnosis platform

通过人工智能算法提供诊断建议，支持图像分类、图像分割、目标识别等，可对疾病进行筛查及诊断，具有疾病诊断或诊断分级功能的信息系统。

3.3

医疗器械 medical device

医疗器械，是指直接或者间接用于人体的仪器、设备、器具、体外诊断试剂及校准物、材料以及其他类似或者相关的物品，包括所需要的计算机软件；其效用主要通过物理等方式获得，不是通过药理学、免疫学或者代谢的方式获得，或者虽然有这些方式参与但是只起辅助作用。

3.4

测试数据集 testing set

用于测试、评估人工智能算法/模型的数据集，类标记对算法来说未知，且测试数据与训练数据、验证数据无交集。

3.5

混淆矩阵 confusion matrix

分类任务中用于快速观察相关数量的矩阵。

3.6

准确率 accuracy

对于给定的数据集，正确预测的样本占所有样本的比率。

3.7

假阴性率 false negative rate

对于给定的数据集，错误预测为阴性病例样本占所有预测为阴性病例样本的比率。

3.8

特异度 specificity

对于给定的数据集，阴性病例样本被正确分类的比率。

3.9

真阳性率 true positive rate

对于给定的数据集，阳性病例样本被正确分类的比率，也被称为召回率、真阳率、灵敏度。

3.10

感兴趣区域 region of interest (ROI)

机器视觉、图像处理中，从被处理的图像以方框、圆、椭圆、不规则多边形等方式勾勒出需要处理的区域，称为感兴趣区域。

4 缩略语

CT：计算机断层扫描（Computed Tomography）
DICOM：医学数字成像和通信（Digital Imaging and Communications in Medicine）
JPEG：连续色调静态图像压缩标准/联合图像专家组（Joint Photographic Experts Group）
MR：磁共振（Magnetic Resonance）
BMP：位图（Bitmap）
TIFF：标签图像文件格式（Tagged Image File Format）
RF：射频（Radio Frequency）
PET-CT：正电子发射断层扫描成像（Positron Emission Tomography－Computed Tomography）

5 质量评估内容

5.1 质量评估内容框架

本文件从智慧医疗影像辅助诊断平台的质量特性出发，结合用户的实际需要，给出了智慧医疗影像辅助诊断平台质量评估内容及评估方法，其内容如图1所示，包含产品文档、功能适合性、功能正确性、性能效率、可靠性、易用性、可移植性、安全性、数据有效性共9个类别，每个类别包含评估的具体内容/指标。

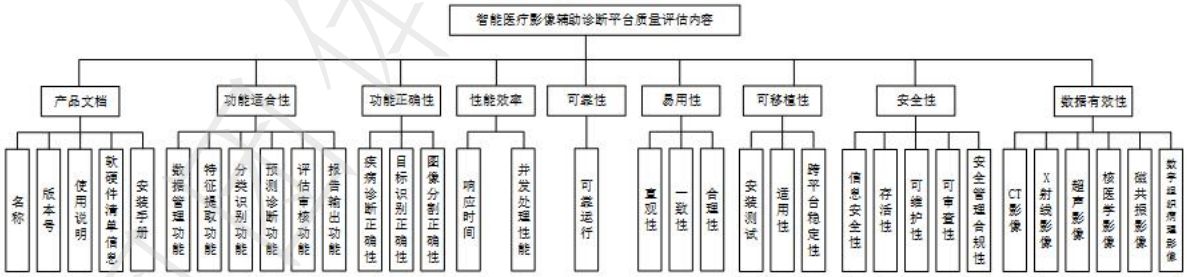

图 1 智慧医疗影像辅助诊断平台质量评估内容框架

5.2 产品文档

5.2.1 名称

智慧医疗影像辅助诊断平台的产品文档应对平台名称进行说明。

5.2.2 版本号

智慧医疗影像辅助诊断平台的产品文档应对平台的软件版本号进行说明。

5.2.3 使用说明

智慧医疗影像辅助诊断平台的产品文档应有完整的使用说明文档或用户手册。

5.2.4 软硬件清单信息

智慧医疗影像辅助诊断平台的产品文档应提供平台涉及的适配软硬件信息清单，包括操作系统、浏览器、服务器硬件、内存等情况。

5.2.5 安装手册

智慧医疗影像辅助诊断平台的产品文档应具有完整的安装手册。

5.3 功能适合性

5.3.1 数据管理功能

平台应支持异构数据的上传、查询、修改、删除、备份、恢复、下载等操作，支持的数据类型宜包括但不限于CT、MR、PET-CT、X射线、超声等医疗影像，支持的文件格式宜包括但不限于DICOM、JPEG/JPG/JPE、TIFF/TIF、BMP等。

5.3.2 特征提取功能

平台支持的特征提取类型宜包括但不限于影像中器官、病灶、ROI区域的颜色、纹理、形状等特征。

5.3.3 分类识别功能

平台宜具有分类、识别或分割医疗影像的能力，包括：

a) 支持手动或自动对器官、疑似病灶或其他感兴趣区域（如邻近脏器、淋巴结、大血管等）进行标记；
b) 支持一种或多种的影像分割，对器官、病灶、ROI 区域进行自动、半自动化分割，达到区分病灶或器官的目的，为定性或定量辅助分析提供基础；
c) 分割处理的输入为经过预处理后的影像或者影像序列，经过处理后的输出为分割后图像或者记录分割点信息的数据结构，标记处理后支持保存原始的输入影像。

5.3.4 预测诊断功能

平台宜具有评估预测分析的能力，包括：

a) 标示疑似病灶的位置；
b) 计算疑似病灶的大小或体积；
c) 分析疑似病灶的性质，必要时给出分级诊断；
d) 计算预测结果的置信度。

5.3.5 评估审核功能

平台宜支持医疗专家对预测诊断模块的输出结果进行审核、修订和确认，包括：

a) 支持医生对模型输出结果（感兴趣区）、文字诊断说明进行修改和确认，并保留修改日志；
b) 支持将医疗专家的修订结果在经过质量评估审核后，纳入标注数据库，并反馈至模型评估流程，支持按时间阶段提取；
c) 审核过程中，支持多位影像领域知识的专家对医生诊断结果进行审核、确认。

5.3.6 报告输出功能

平台宜支持在医疗专家的最终决策和确认下，出具诊断报告，包括：

a) 支持自动生成诊断报告；
b) 支持医疗专家修改、确认、发布；
c) 报告输出内容宜包括：门诊信息、患者信息、指标说明或影像描述、诊断结论、建议及相关医疗人员说明等；
d) 诊断报告术语应符合标准医学术语规范，内容输出宜采用结构化模板。

5.4 功能正确性

5.4.1 疾病诊断正确性

平台的疾病诊断正确性评估宜选用的评估指标包括但不限于精准度、准确率、假阴性率、敏感度、特异度、召回率和F1值等，具体评估方法（/要求）详见附录A.1.1。

5.4.2 目标识别正确性

平台的目标识别正确性评估宜选用的评估指标包括但不限于准确率、召回率、F1值、平均精度、平均精度均值等，具体评估方法（/要求）详见附录A.1.2。

5.4.3 图像分割正确性

平台的图像分割正确性评估宜选用的评估指标包括但不限于准确率、召回率、F1值、图像分割的交并比、平均交并比等，具体评估方法（/要求）详见附录A.1.3。

5.5 性能效率

5.5.1 响应时间

平台质量评估中的性能效率评估指标应包括响应时间,即平台对数据样本进行运算得到结果所需要的平均时间，具体评估方法详见附录A.2。

5.5.2 并发处理性能

平台质量评估中的性能效率评估指标应包含并发处理性能,即平台同时能处理分析的数据样本数量。

5.6 可靠性

平台的可靠性满足以下要求:

a) 平台在用户相关文档描述的限制范围内使用时，软件不应丢失数据；
b) 平台应识别违反句法条件的输入，并且不应作为许可的输入加以处理；
c) 平台应具有从严重错误中恢复的能力；
d) 平台负载运行情况下，在指定的一段时间内应保持正常稳定持续运行。

5.7 易用性

5.7.1 直观性

平台的UI界面直观性应满足以下要求：

a) 操作界面状态显示明确清晰：包括图标控件文字状态、界面说明、操作状态等；
b) 操作反馈清晰：对重要的操作有明确的状态提示信息，并允许用户取消操作。

5.7.2 一致性

平台的操作界面菜单和热键应符合惯性操作、市场主流，术语和命名运行应符合相关法规标准。

5.7.3 合理性

平台的UI界面合理性应满足以下要求：

a) 界面设计符合规范性、合理性原则，各个版面布局、视觉风格、图标空间、字体、导航应保持一致；
b) 菜单结构、信息结构清晰，逻辑明确，主次突出。

5.8 可移植性

5.8.1 安装测试

平台在正常情况下的多种条件或设置，可进行安装以及正常卸载。如：多种软硬件环境的安装测试、安装顺序测试、安装启动测试、修复安装测试和卸载测试、更新包测试等。

5.8.2 适用性

平台能够在所指定的目标环境（硬件、软件、中间件、操作系统等）中正确地运行。

5.8.3 跨平台稳定性

平台能够在软硬件环境发生一定变化的情况下，保持正常运行且正确完成任务。

5.9 安全性

5.9.1 信息安全性

5.9.1.1 保密性

平台根据授权类型与授权级别来区分访问数据的权限，用户只有在被授权时才能正常访问数据信息。在必要的场合下，平台宜采用加密算法对用户的数据信息进行加密存储或传输，具备数据加密保护机制。

对数据中的个人敏感信息应按照各医疗机构的规定进行脱敏，对需要使用的敏感信息应采取最小必要原则。

5.9.1.2 完整性

平台应妥善存储数据，且存储在数据库中的所有数据值均正确。防止未授权访问数据信息，避免程序或数据被篡改。

5.9.1.3 不可抵赖性

在平台进行信息交互的过程中，所有参与者都不可否认或抵赖曾经完成的操作和承诺。

5.9.2 存活性

平台具有一定的容错性，尽管存在硬件或软件的某些故障，其系统、产品或组件的运行应符合预期的效果。

当发生中断或失效时，平台能够在一定程度上恢复直接受影响的数据并重建期望的系统状态。

平台具有一定抗风险能力，在受到攻击时，平台能够尽快做出反应，防止信息泄露，并及时提供必要的服务，继续履行其任务。

5.9.3 可维护性

平台应具备可维护性，符合GB/T 25000.51-2016 中相关要求。

平台系统或组件能够被维护人员检查与修改，包括平台的安装更新、软硬件环境的适应性调整。

平台应支持运行状态监控，包括平台系统运行时动态性能信息、意外失效和重要条件的信息、运行指示器（如日志、警告屏）的信息和处理本地数据的信息等。

5.9.4 可审查性

平台应具备可审查性，能够对操作以及操作人员信息进行审查，实体的活动能够被唯一地追溯。

平台具备审计功能，能对管理员登录、系统崩溃等敏感事件严格鉴权，同时将操作信息进行记录，且存留日志信息不少于6个月。

5.9.5 安全管理合规性

平台运行、使用应恪守医学科研伦理，按照现行要求进行伦理审评和备案。

医学影像信息的采集、标注、使用、以及共享等应获得医疗机构的伦理审批。以知情同意书形式获得个人信息主体的知情权与授权，或伦理机构同意免除知情通知；数据操作严格遵守ISO 14155:2020的要求。

5.10 数据有效性

5.10.1 CT 影像

5.10.1.1 应用场景

平台支持的CT影像应用场景可包括但不限于：

a) 全身各部位新生物的检出及其定位、定性诊断，肿瘤分布范围，浸润和转移及 CT 引导下的活检；
b) 全身各部位炎症检出及其定位、定性诊断和范围大小的确定；
c) 全身各部位大血管病变的检出和定性诊断，冠状动脉病变的范围、程度和斑块性质的确定；

d) 重要脏器外伤出血的定量及定性，多种外伤异物的定位；
e) 某些器官部位的钙化或结石检出；
f) 脏器变性（如肝脂肪变性）或先天异常。

5.10.1.2 数据格式与存储

平台支持的数据文件格式应包括：DICOM、JPEG、BMP等。

5.10.1.3 影像质量

适用于平台的CT影像质量应至少达到隐约可见要求，即器官和结构在检查范围内可观察到，但细节未显示。

若用户对影像质量有高要求，则需达到可见或清晰显示要求，即解剖结构细节可见，但不能清晰辩认，也称细节显示；或者解剖细节清晰辩认，即细节清晰。

5.10.2 X射线影像

5.10.2.1 应用场景

平台支持的X射线影像场景可包括但不限于：
a) 数字X线普通摄影，
b) 数字乳腺摄影，
c) 数字减影血管造影，
d) 数字胃肠摄影。

5.10.2.2 数据格式与存储

平台支持的数据文件格式应包括：DICOM、JPEG、BMP等。如有RF信号数据，亦可保存为二进制RF格式。

5.10.2.3 影像质量

适用于平台的X射线影像质量要求应至少达到隐约可见要求,即解剖学结构或/和病变特征等细节可观察到，但细节没有完全显示，只特征可见。

若用户对影像质量有高要求，则需达到可见或清晰显示要求，即解剖学结构或/和病变特征细节可清晰辨认，也称细节清晰；或者解剖学结构或/和病变特征等细节可见，但不能清晰辨认，也称细节显示。

5.10.3 超声影像

5.10.3.1 应用场景

平台支持的超声影像应用场景可包括但不限于：
a) 浅表器官及血管病变（临床较多为乳腺、甲状腺、眼、动静脉血栓等）；
b) 超声引导下穿刺；
c) 腹部脏器病变及实体肿瘤（临床较多为肝脏、胰腺、肾等）；
d) 妇科检查（如子宫、卵巢）。

5.10.3.2 数据格式与存储

平台支持的数据文件格式应包括：DICOM、JPEG、BMP等。如有RF信号数据，亦可支持二进制RF格式。

5.10.3.3 影像质量

适用于平台的超声影像质量应达到但不限于以下要求：
a) 图像尺寸不小于800像素×600像素；
b) 除疾病性质导致的特殊情况外，图像清晰、均匀，超声切面标准，图无斑点、雪花细粒、网纹等干扰。

5.10.4 核医学影像

5.10.4.1 应用场景

平台支持的核医学影像应用场景可包括但不限于：

a) 胸腹部脏器实体肿瘤（临床较多为肺、结直肠、前列腺等）；
b) 神经系统显像，主要为大脑；
c) 部分特殊器官显像，如心脏。

5.10.4.2 数据格式与存储

平台支持的数据文件格式应包括：DICOM、JPEG、BMP等。

5.10.4.3 影像质量

适用于平台的核医学影像质量应达到但不限于以下要求：

a) 图像尺寸不小于 800 像素×600 像素；
b) 图像清晰、均匀，超声切面标准，图无斑点、雪花细粒、网纹等干扰。

5.10.5 磁共振影像

5.10.5.1 应用场景

平台支持的磁共振（MR）影像应用场景可包括但不限于：

a) 中枢及外周神经系统病变及肿瘤鉴别和分级；
b) 心血管系统病变检出及心功能评估；
c) 肝胆胰病变检出及肿瘤鉴别；
d) 泌尿系统及前列腺病变检出及肿瘤分级；
e) 胃、结直肠病变检出和鉴别；
f) 骨组织病变及肿瘤鉴别；
g) 关节、软骨损伤评估；
h) 乳腺肿瘤鉴别与分级；
i) 各类外伤导致的器官、组织损伤评估。

5.10.5.2 数据格式与存储

平台支持的数据文件格式应包括：DICOM、JPEG、BMP等。

5.10.5.3 影像质量

适用于平台的磁共振影像质量应达到但不限于以下要求：

a) 常规图像无明显畸变，弥散加权图像畸变在正常范围内；
b) 图像信号均匀，无明显信噪比异常，无明显射频干扰及其他各种形式的伪影；
c) 图像无明显运动伪影。

5.10.6 数字组织病理影像

5.10.6.1 应用场景

平台支持的数字组织病理影像应用场景可包括但不限于：

a) 常规组织学诊断，
b) 冰冻诊断，
c) 免疫组化，
d) 免疫荧光，
e) 细胞学诊断。

5.10.6.2 数据格式与存储

平台支持的数据文件格式应包括：Generic tiled TIFF、JPEG、BMP、JPEG2000等。

5.10.6.3 影像质量

适用于平台的数字组织病理影像质量应达到但不限于以下要求：

a) 切片完整情况：切片完整、无破损情况；
b) 切片染色情况：染色保持鲜亮，视觉感知明显，能清晰辨识细胞大小及形状；
c) 细胞以及核边界清晰；
d) 包含细胞分布区域95%以上；
e) 未引入涂片不包含的图像噪声，无外部噪声影响阅片诊断。

6 测试评估方法

6.1 测试评估流程

智慧医疗影像辅助诊断平台质量评估流程如图2所示，主要包括建立评估模型、构建测试环境、生成测试数据集、进行平台评估、获取测试结果等步骤。

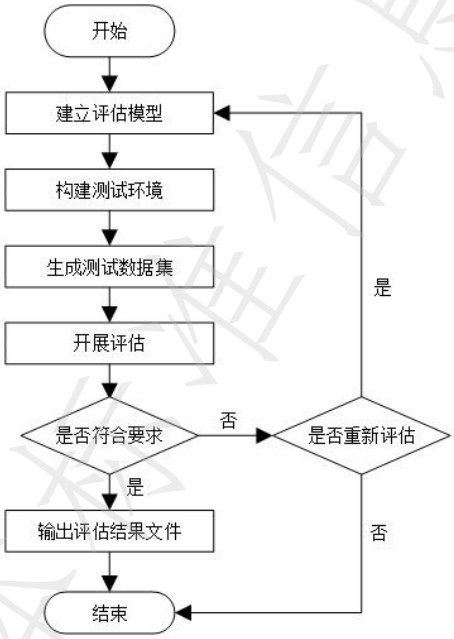

图 2 智慧医疗影像辅助诊断平台质量评估流程

6.2 建立评估模型

根据本文件5.2-5.10的一系列评估要求或评估指标，建立评估模型，给定评估指标通过准则。平台进行测试评估时，宜选取较为全面的评估指标。

6.3 构建测试环境

根据平台所需的软硬件要求，构建相应的平台测试环境，测试环境应符合待测产品的产品文档指明的运行环境需求。若无法构建与实际使用场景一致的测试环境，则需要另外说明差异点以及差异点可能会带来的影响。

当实验室无法满足检测所需的环境时，由送测方协助提供特定设备或检测环境。

6.4 生成测试数据集

按照实际场景需求，生成、构建覆盖评估指标需求的测试数据集用以开展测试。对测试数据集的要求包括：

a) 数据集覆盖所选评估指标需要的医疗影像；
b) 数据集若包含标签，则标注的标签应正确无误；
c) 数据集描述应包括数据来源、数据分布等信息；

d) 对于数据质量要求较高的任务场景，数据来源宜采用来自多家医疗机构的、某个长跨度时间段的临床真实数据，以保证数据的多样性；
e) 当数据涉及隐私保护的情况下，应收集能够证明数据唯一性的辅助信息，如样本 ID、标签、生成时间等不涉及商业秘密的信息。收集数据时，还应注意数据的格式信息与存储信息；
f) 根据平台实际业务的数据需求，必要时可对数据进行处理并生成测试数据集，处理方式可为增加噪声、增加扰动。

6.5 开展评估

在构建的环境下，根据选择的评估指标逐一进行测试，测试结果应符合本文件要求。详尽记录评估过程所产生的数据，包括但不限于样本的ID、预测结果、置信度、图像分割中分割区域的点值、图像识别中的预测框等。其中，评估过程中涉及的计算公式参考附录A。

6.6 评估报告

衡量评估结果是否满足6.2的要求，并依据建立的评估模型得出评估结论，判断评估活动的结果。评估报告宜采用表格或图展示统计信息，如混淆矩阵、ROC曲线等。

附 录 A
（规范性）
功能正确性及性能效率评估方法

A.1 功能正确性评估方法

A.1.1 疾病诊断正确性

根据实际的应用场景选择任务相关的基本指标，用于评估算法模型完成功能的能力，如分类任务中的精准度、准确率、假阴性率、敏感度、特异度等，评估方法如下：

混淆矩阵是指分类任务中用于快速观察相关数量的矩阵，矩阵大小为。

以较常见的疾病诊断二分类任务为例，混淆矩阵按照公式（A.1）计算。

$$M=\begin{array}{c|cc} \text{真实 \ 预测} & \text{正例} & \text{负例} \\ \text{正例} & TP & FN \\ \text{负例} & FP & TN \end{array} \quad \cdots\cdots (A.1)$$

式中：

TP（True Positive）——真正例；

TN（True Nagative）——真负例；

FP（False Positive）——假正例；

FN（False Negative）——假负例。

准确率ACC，关注被正确诊断的样本数量，取值范围为[0,1]。当正确诊断的正例、负例越多，则准确率越高。但在样本不平衡的情况下，不宜使用准确率作为评估指标。

准确率ACC按照公式（A.2）计算。

$$ACC=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \quad \cdots\cdots (A.2)$$

式中：

TP（True Positive）——真正例；

TN（True Nagative）——真负例；

FP（False Positive）——假正例；

FN（False Negative）——假负例。

召回率REC，也称为灵敏度、真正率TPR，关注真实正例的预测情况，计算正确分类的正样本与所有分类为正的样本之比。召回率取值范围为[0,1]，当遗漏的真正例越少，取值越接近于1，则召回率越高。

召回率REC按照公式（A.3）计算。

$$REC=\frac{TP}{TP+FN}\times 100\% \quad \cdots\cdots (A.3)$$

式中：

TP（True Positive）——真正例；

FN（False Negative）——假负例。

特异度SPEC，与召回率相反，关注真实负例的预测情况，计算正确分类的负样本与所有分类为负的样本之比。特异度取值范围为。当遗漏的真负例越少，取值越接近于1，则特异度越高。

特异度SPEC按照公式（A.4）计算。

$$SPEC=\frac{TN}{TN+FP}\times 100\% \quad \cdots\cdots (A.4)$$

式中：

TN（True Negative）——真负例；

FP（False Positive）——假正例。

精度PREC，关注模型的预测结果，计算为正确分类的样本与分配给该类的所有样本之间的比率。其特点可以与召回率进行对比，区别在于精度更关注模型预测结果，而召回率更关注于样本身被预测的情况。精度取值范围为[0,1]，其中1表示该类中的所有样本都正确预测，0表示该类中没有正确预测。

精度PREC按照公式（A.5）计算。

$$\mathrm{PREC}=\frac{\mathrm{TC}}{\mathrm{TC+FC}} \quad \text{(A.5)}$$

式中：

C——类别，二分类时可以由P（正）或者N（负）代替。

假阴性率FNR，关注误诊的阳性病例的数量，错误预测为负例的数量与所有分类为正的样本之比。取值范围为[0,1]，实际中期望假阴性率越趋与0越好。

假阴性率FNR按照公式（A.6）计算。

$$\mathrm{FNR}=\frac{\mathrm{FN}}{\mathrm{TP+FN}}\times 100\% \quad \text{(A.6)}$$

式中：

TP（True Positive）——真正例；

FN（False Negative）——假负例。

F1值，是准确率和召回率的调和平均值，F1给予较小值更高的权重。F1值限制为[0,1]，其中1表示最大精度和召回值，0表示零精度和/或召回。

F1值按照公式（A.7）计算。

$$\mathrm{F1}=2\times\frac{\mathrm{precision}\times\mathrm{recall}}{\mathrm{precision+recall}}=\frac{2\times\mathrm{TP}}{2\times\mathrm{TP+FP+FN}} \quad \text{(A.7)}$$

式中：

TP（True Positive）——真正例；

FP（False Positive）——假正例；

FN（False Negative）——假负例。

但值得注意的是，该指标在类之间不是对称的，取决于哪个类被定义为正类和负类。例如，在一个样本的正类数较多且分类器偏向于正类的情况下，与TP成正比的F1分数会很高。在不改变类分布的条件下，重新定义类标签以使负类占多数并且分类器偏向负类将使得F1值降低。

假阳性率FPR值，与真阳性率TPR相对，关注假正例的预测情况，计算错误分类的正样本与所有分类为负的样本之比。假阳性率取值范围为[0,1]，值越小代表性能越好。

真阳性率TPR按照公式（A.8）计算。

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP+TN}}\times 100\% \quad \text{(A.8)}$$

式中：

FP（False Positive）——假正例；

TN（True Negative）——真负例。

AUC值，在给定测试环境下，ROC曲线（Receiver Operating Characteristic Curve）表示接收者操作特征曲线，AUC值即ROC曲线下的面积，其取值一般介于0.5和1之间，作为数值可以直观地评价分类器的好坏，值越大代表分类效果越好。

结合前文的TPR值和FPR值，AUC值可按照公式（A.9）计算。

$$\mathrm{AUC}=\int TPR\ d(FPR) \quad \text{(A.9)}$$

式中：

TPR（True Positive Rate）——真阳性率；

FPR（False Positive Rate）——假阳性率。

A.1.2 目标识别正确性

目标识别任务的评价指标可依照公式（A.1）中的混淆矩阵展开，以目标识别任务真实框（Ground Truth Bounding Box）作为基准，TP的根据是否达到IoU（重叠度）阈值进行计算。以IoU阈值为0.5举例，含义如下：

a) 当预测标签为目标（正）且预测框的时，则认为该预测框为正确预测结果，TP 的计数值加 1；

b) 当预测标签为目标（正）且预测框的时，则认为该预测框为错误预测结果，FP 的计数值加 1；

c) 预测标签为背景（负）且预测框的时，则认为该预测框为错误预测结果，FN 的计数值加 1；

d) 预测标签为背景（负）且预测框的时，则认为该预测框为正确预测结果，TN 的计数值加 1。

重叠度（交并比）IoU（Intersection over Union），也称为交并比，检测结果的预测框与样本标注的矩形框的交集与并集的比值。

重叠度（交并比）IoU按照公式（A.10）计算。

$$IoU = \frac{预测框 \cap 真实框}{预测框 \cup 真实框} \quad (A.10)$$

在以上说明的基础上，目标识别正确性的评价指标如准确率、召回率、精度和F1值等，可按照A.1.1中的说明进行计算。

平均精度AP（Average Precision），是指设定IoU值之后不同置信度下的精度加权平均值，每一项的权重是对应置信度的召回率，其计算过程等同于计算Precision-Recall曲线下方的面积。

计算AP的步骤如下：

a) 使用模型生成预测分数（predict score）。

b) 将预测分数转换为类标签。

c) 计算混淆矩阵中的 TP、FP、TN、FN。

d) 计算精度和召回率。

e) 计算精确召回曲线下的面积。

f) 计算平均精度。

平均精度AP按照公式（A.11）计算。

$$P = \sum_{confidence_i} Precision_i \times Recall_i \quad (A.11)$$

平均精度均值mAP（mean Average Precision）的计算方法是指在有多个标签类的情况下，找到每个类的平均精度(AP)，然后对多个类进行平均。

平均精度均值mAP按照公式（A.12）计算。

$$mAP = \frac{1}{N}\sum_{N}^{i} AP_i \quad (A.12)$$

式中：

N——标签类的数量。

A.1.3 图像分割正确性

图像分割的交并比IoU（Intersection over Union）可按照公式（A.10）进行计算。

平均交并比MIoU（Mean Intersection over Union）计算方法为在每个类上计算IoU之后平均，计算方式如公式（A.13）所示。

$$mIoU = \frac{1}{k}\sum_{i=1}^{k} IoU_k \quad (A.13)$$

式中：

K——样本中类别数量。

在以上说明的基础上，图像分割正确性的评价指标如准确率、召回率、精度和F1值等，可按照A.1.1中的说明进行计算。

A.2 性能效率评估方法

响应时间RT（Response Time），在给定测试环境下，辅助诊断平台对数据样本进行测试运算得到结果所需要的平均时间，假设给定的样本数据的数量为n，和分别表示样本输入的时间和样本结果输出的时间。

响应时间RT计算方式如公式（A.14）所示。

$$RT = \frac{1}{N}\sum_{1}^{N}(T_e - T_b) \quad (A.14)$$

式中：

N——样本数据的数量；

T_e——样本结果输出的时间；

T_b——样本输入的时间。

响应时间在保证准确率的同时，应有较短的响应时间。

参 考 文 献

[1] GB/T 20271—2006 信息安全技术 信息系统通用安全技术要求

[2] GB/T 25000.10—2016 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第10部分：系统与软件质量模型

[3] GB/T 25000.51—2016 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第51部分：就绪可用软件产品（RUSP）的质量要求和测试细则

[4] GB/T 28452—2012 信息安全技术 应用软件系统通用安全技术要求

[5] GB/T 35295—2017 信息技术 大数据 术语

[6] T/CESA 1026-2018 人工智能 深度学习算法评估规范

[7] T/CESA 1037—2019 信息技术 人工智能 面向机器学习的系统框架和功能要求

[8] T/CESA 1109—2020 智慧医疗影像辅助诊断系统技术要求和测试评价方法

[9] ISO/IEC 2382:2015, Information technology—Vocabulary

[10] ISO 20916:2019 In vitro diagnostic medical devices—Clinical performance studies using specimens from human subjects-good study practice

[11] ISO/TR 24291:2021 Health informatics—Applications of machine learning technologies in imaging and other medical applications

[12] ISO/IEC/IEEE 24765:2017, Systems and software engineering—Vocabulary

[13] ISO/IEC/TR 29119-11, Software and systems engineering—Software testing—Part 11: Guidelines on the testing of AI-based systems

[14] IEC 81001-5-1: The standard for secure health software

[15] Beam AL, Kohane IS, Big data and machine learning in health care. JAMA 2018;319:1317–1318.

[16] Benjamens S., Dhunnoo P., Meskó B., The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digit. Med. 2020;3, 118.

[17] The Lancet. Artificial intelligence in health care: within touching distance. Lancet 2018;390:2739

[18] DTXX—2017—11436 《医疗器械分类目录》

[19] FGWJ—2021—10001 《医疗器械监督管理条例》