ICS 11.040.99
CCS C 30

中华人民共和国国家标准

GB/T XXXXX—XXXX

人工智能医疗器械 心电辅助分析软件 算法性能测试方法

Artificial intelligence medical device - Computer assisted analysis software for electrocardiogram —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

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。

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：

引　言

随着人工智能技术在医疗器械领域的快速渗透，人工智能心电辅助分析软件作为典型的人工智能医疗器械产品，已广泛应用于临床心电诊断场景，为医务人员提供心律失常筛查、心肌缺血预警等辅助决策支持，对提升基层医疗机构心电诊断能力、提高诊疗效率、降低漏诊误诊风险具有重要意义。

然而，由于心电信号具有个体差异大、易受干扰、病理类型复杂等特点，加之不同厂商的人工智能算法模型架构、训练数据来源、特征提取方式存在显著差异，导致当前市场上同类产品的算法性能缺乏统一、科学、规范的测试评价依据。上述问题不仅影响了临床用户对产品性能的准确认知，也给监管部门的产品上市审评、上市后监督带来挑战，难以有效保障医疗质量和患者安全。

本标准规定了人工智能心电辅助分析软件算法性能的测试方法，适用于该类软件在产品研发、注册检验、上市后评价等阶段的算法性能测试，旨在推动人工智能心电辅助分析软件行业的规范化发展，进一步推动人工智能技术与心电诊疗深度融合，为提升心血管疾病诊断效率、降低诊疗风险提供技术支撑，助力健康中国建设。

人工智能医疗器械 心电辅助分析软件 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心电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的心电辅助分析软件产品。

本标准不适用于心电图采集硬件设备或系统、用于处理非ECG信号（如光电容积描记（PPG）信号）的AI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。

GB/T 9386-2016 计算机软件测试编制规范

GB/T 11457-2006 信息技术软件工程术语

GB/T 35295-2017 信息技术大数据术语

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

ISO/IEC TS 4213 信息技术 人工智能 机器学习分类性能评估

IEC 63521 机器学习医疗器械 性能评估过程

3 术语和定义

GB/T 35295-2017、GB/T 11457-2006和YY/T 1833.1-2022界定的以及下列术语和定义适用于本文件。

3.1

心电图诊断设备人工智能辅助判断 ECG diagnostic equipment artificial-intelligence-aided judgment

通过诊断设备提取人体心电信号中QRS等病理特征波，再经过基于人工智能的计算机技术分析，由AI给出判断结果，并筛选出可疑的病变如早搏、心动过速等病理，辅助医生更快速，准确的做出心电病理诊断。

3.2

心电图生理信号标注 ECG physiological signal annotation

对心电图的心拍标记、分类等信息的物理量进行分析，建立外部知识的过程。

3.3

测试计划 test plan

描述预定测试活动的范围、方法、资源和进度的一种文档。它确定测试项、要测试的特征、测试任务、执行每一任务的人员以及需要应急对策的任何风险。

3.4

重复筛查 repeat screening

以一定周期重复进行的筛查。

3.5

压力测试 stress test

使用具有挑战性的用例或测试集开展测试的过程。

4 缩略语

AI：人工智能(Artificial Intelligence)
ECG：心电图（electrocardiogram）
TCA：总体分类精度(Total classification accuracy）
PPV：阳性预测值(Positive predictive value)
PGD：白盒攻击(Projected Gradient Descent)

5 概述

人工智能医疗器械相关标准的一般原则适用于本文件所涉及的心电设备，本文件提出了用于分析心电设备的人工智能心电辅助分析软件算法的性能测试的指标与要求。

算法性能测试是心电信号辅助分析软件验证与确认的重要环节，一般基于测试集对算法进行评估，对算法输出结果和参考标准进行定量比较，实现假阳性与假阴性、重复性与再现性、鲁棒性/健壮性、效率等具体指标的评估。

本文件描述了心电辅助分析软件算法独立性能测试的方法，涵盖测试流程、算法性能测试方法、算法质量特性与测试方法，同时考虑人工智能医疗器械的特殊质量特性，例如泛化性、鲁棒性、重复性、一致性和效率。

6 测试前准备

6.1 测试环境

6.1.1 硬件环境

计算设备： 测试应在一台或多台具有代表性的硬件平台上进行。硬件配置（如CPU型号、核心数、内存容量、GPU型号与显存等）应明确记录，并至少满足软件用户文档集规定的最低和推荐配置。测试应覆盖典型配置（如主流商用电脑）和高性能配置（如工作站）。

心电图采集设备： 若测试涉及与心电采集设备的连接（如动态心电记录仪、静息心电采集盒），应使用已注册、临床常规使用的设备型号，并记录设备型号、固件版本。

外围设备： 包括显示器（分辨率、尺寸）、存储设备等，应满足软件正常运行的基本要求。

6.1.2 软件环境

操作系统： 测试应在软件声称支持的所有操作系统版本（例如Windows 特定版本，特定Linux发行版或国产操作系统等）上进行，并记录确切的OS版本号和补丁级别。

支撑软件与依赖库： 记录所有必需的第三方软件运行环境、依赖库及其版本号（如 数据库、中间件等）。

被测软件版本：应清晰记录被测软件的唯一标识,包括:软件名称、版本号、构建号(Build Number)、哈希值（如SHA-256）。测试版本应为最终申报的、不可更改的发布版本。

安全软件： 记录测试环境中安装的杀毒软件、防火墙等安全软件的配置情况，以避免其对软件性能产生干扰。

6.1.3 网络环境

如软件为本地部署，则测试应在断网环境下进行，以排除未知更新或网络请求干扰。如软件为云端部署，则网络带宽与延迟、稳定性应满足测试所需性能。

6.2 测试资源

6.2.1 数据集

测试集是软件算法测试的基础，为更好开展测试验证工作，测试集需满足如下要求：

a) 测试集的质量应满足心电图诊断设备人工智能辅助判断数据集要求，并独立于算法研发、训练、调优过程，保证封闭性和安全性；

b） 测试集重复使用率需受到限制，以防止被测试软件在测试过程中进行学习；
c） 测试集的来源需使用已上市的心电产品，如产品未上市，则需进行评估后使用；
d） 测试集的样本总量应不低于单次测试样本总量的 n 倍（n 根据实际测试场景确定），并涵盖人口统计、病种等信息。

注：附录A给出测试集描述的示例。

6.2.2 样本量与样本分布

测试数据集的样本量与分布应经过科学设计，以确保性能评估结果具有足够的统计置信度，并能代表软件在目标患者人群和预期使用环境中的真实性能。

a） 应采用灵敏度计算单次测试中阳性样本的样本量，用特异度计算单次测试中阴性样本的样本量；
b） 应根据用途，对测试集的样本分布进行分析，包括人口统计学、用途、地域、疾病分布等；
c） 特殊情况处理：对于某些罕见疾病，若无法满足最低样本量要求，应收集所有可获得的经确认的病例进行测试，并在报告中明确说明其局限性，指出性能评估的不确定性。

6.2.3 扩增数据

为评估算法在面对预期范围内的信号变化和干扰时的性能保持能力，可采用白盒或黑盒方式在独立测试集的基础上引入受控的、模拟真实世界情况的数据扩增来实施。

数据扩增应考虑如下要求：

a） 白盒扩增方式：其内部算法内部结构、特征提取逻辑、决策机制均为可理解；
b） 黑盒扩增方式：其算法内部结构、特征提取逻辑以及决策机制毫不知情，仅关注输入输出；
c） 数据核对：对于扩增后的仿真数据，应于真实数据进行比较论证，必要时可进行抽样检测；
d） 扩增数据集要求应符合 6.2.1 要求，其使用记录应于真实世界数据严格区分并独立记录。

6.3 测试流程要求

6.3.1 测试准备

测试流程应遵循完整的闭环管理，包括测试准备、测试执行、结果收集与分析、问题管理等环节，以确保测试活动的有序性和结果的可追溯性。具体各步骤要求如下：

a） 制定测试方案和计划，并开展评审工作；
b） 根据评测产品，搭建测试资源环境；
c） 根据数据集要求，准备测试数据集，并转化为测试产品所要求的格式，确保软件产品正常加载数据集。

6.3.2 测试执行

测试执行包括：

a） 预测试：在正式测试之前，可应用小部分代表性测试数据开展预测试，以验证测试流程的完整性，确认软件正常启动、加载数据并输出结果；
b） 正式测试：测试过程中应尽可能通过自动化脚本或工具执行，降低人为操作误差，并对每一条输入的测试数据，必须完整记录软件的原始输出结果，同时确保输入数据与输出数据有唯一标识绑定。

6.3.3 结果收集与处理

应收集并备份所有测试用例的软件原始输出文件、系统日志及性能监控日志，并通过自动化工具与输出结果与金标准进行对比，其中对比应支持按软件功能、详情数据多种维度开展。

6.3.4 问题管理与回归

问题管理和回归要求：

a） 对于所有假阳性、假阴性及严重偏差的案例，应作为“问题”被详细记录，并组织临床专家和算法工程师对问题进行复核，明确问题原因；

b) 对软件在测试后可及时修复的问题，应开展回归测试，回归测试流程应与正式流程相同，且覆盖所有已修复的问题用例。

6.4 测试报告

测试报告应包含如下信息：

a) 测试环境：应包含硬件环境、软件环境、网络环境等；
b) 测试对象：应包含被测试产品的名称、版本号、型号、部署模式、固件号、制造商等；
c) 测试资源：应包含数据集、对比产品、标准品等；
d) 测试方法与判断标准：应包含测试项目、方法（功能性测试、非功能性测试）等；
e) 测试结果：应采用多种形式记录测试项目的实际结果；
f) 测试结论：应对本次测试结果进行评判，是否通过本次测试，如未通过应明确项目及原因；
g) 附录：其他相关文件。

7 算法性能测试方法

7.1 针对算法应用场景的测试方法

7.1.1 心电图特征波形检测与分类

7.1.1.1 标记与匹配

针对心电特征波形进行检测识别、分类，如P波、QRS波和T波。波形检测通常为波形峰值点定位，即识别波形的位置。

对于特征波形检测，P波和T波计算算法检测位置与金标准的位置重合度（波形位置距离≤120ms视为匹配）。QRS波计算算法检测位置与金标准的位置重合度（波形位置距离≤100ms视为匹配）。若以金标准的位置进行匹配，则以其为中心，如果匹配窗内有多个待匹配的算法检测结果，则视距离最近的一个算法检测结果为匹配结果；若以算法检测位置进行匹配，则以其为中心，如果匹配窗内有多个待匹配的金标准的位置，则视距离最近的一个金标准的位置为匹配结果。每个金标准的位置和每个算法检测位置，最多均只能被匹配一次。根据匹配结果，每种特征波形的算法检测结果可分为三种：

a) 真阳性，即算法检测结果与金标准匹配，统计匹配波形总数记为 TP;
b) 假阳性，即算法检测结果未能与金标准匹配，统计未匹配波形总数记为 FP;
c) 假阴性，即金标准未被算法检测结果匹配，统计未被匹配的金标准波形总数，记为 FN。

波形分类通常是将波形按照一定的分类标准分为不同类别，如形态分类、心电起源分类等。如：P波/T波按形态分类为：正向、负向、正负双向、负正双向；QRS波按照起源分类：窦性、房性、交界性、室性、起搏等。每种特征波形的算法分类结果（如算法声称可分类为A、B、C），直接与金标准进行比对，结果可分为四种（以待评价分类结果A为例，A定义为阳性，非A（B和C）则定义为阴性）：

a) 真阳性，即金标准结果为 A，算法分类为 A，算法正确分类为阳性，统计此正确分类的波形总数记为 TP;
b) 假阳性，即金标准结果为非 A，算法分类为 A，算法误分类，统计误分类波形总数记为 FP;
c) 假阴性，即金标准结果为 A，算法分类为非 A，算法漏分类，统计漏分类的波形总数记为 FN。
d) 真阴性，即金标准结果为非 A，算法分类为非 A，算法正确分类为阴性，统计此正确分类的波形总数记为 TN。

为方便后续计算各性能指标，针对分类任务，可使用表1所示的混淆矩阵对数据测试结果进行初步分析。

表1 n 分类混淆矩阵

分类		算法分类结果					
		Pred_1	Pred_2	…	…	…	Pred_n
金标准	True_1	$N_{1,1}$	$N_{1,2}$	…	…	…	…
	True_2	…	$N_{2,2}$	…	…	…	…
	…	…	…	…	…	…	…

	…	…	…	…	…	…	…
	…	…	…	…	…	…	…
	True_n	…	…	…	…	…	N_{nn}

注：Pred_x (x=1～n)为算法分类为x类的类别；True_X(x=1～n)为金标准分类为x类的类别；Ni,j (i=1～n, j=1～n) 为金标准的分类结果为i类，被算法分类为j类的个数；n为分类类型个数。二分类的混淆矩阵可简化为表2所示。

表2　二分类混淆矩阵

分类		算法分类	
		阳性	阴性
金标准分类	阳性	TP	FN
	阴性	FP	TN

7.1.1.2　灵敏度与漏诊率

灵敏度（Sensitivity，Sen），即等同于召回率（Recall），本文仅以灵敏度进行说明，其反映了算法对阳性样本的检出能力，同样的，也反映了算法的漏诊率，通用计算公式定义如下：

$$Sen = \frac{TP}{TP + FN} * 100\% \quad \text{……（1）}$$

$$漏诊率 = 1 - Sen \quad \text{……（2）}$$

考虑到心电图的个体差异较大，相同采样时间内，因个体心率不同，所包含的特征波形总数差异较大；同时，心电图存在心律失常时，个体特征波形的分类情况也各不相同。因此，考虑到个体差异，建议同时使用如下两种方法进行灵敏度测量，更全面评估算法的性能，尤其是在个体差异较大或数据分布不平衡的情况下。假设数据集内有N个样本（数据），每个样本（数据）包含的特征波形及类型各不相同：

1）以特征波形为单位，计算所有样本所有特征波形的灵敏度Sen_Gross（以Sen_G标识）

$$Sen\_G = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FN_i)} * 100\% \quad \text{……（3）}$$

其中，i= 1~N，表示N个样本（数据），TPi表示第i个样本的特征波形的真阳性个数，FNi表示第i个样本的特征波形的假阴性个数；

2）以样本（数据）为单位，计算所有样本灵敏度的加权平均值Sen_Average（以Sen_A标识）

$$Sen_i = \frac{TP_i}{TP_i + FN_i} * 100\% \quad \text{……（4）}$$

$$Sen\_A = \frac{\sum_{i=1}^{N} Sen_i}{N} \quad \text{……（5）}$$

其中，N表示样本（数据）总数，Sen_i表示第i个样本(数据)的敏感度。

7.1.1.3　阳性预测值

阳性预测值PPV，即与精确度（查准率）等同，其同时反映了算法的误检和漏检情况，通用计算公式定义如下：

$$PPV = \frac{TP}{TP + FP} * 100\% \quad \text{……（6）}$$

类似Sen，考虑到个体差异，同样建议使用如下两种方法进行阳性预测值测量，以更全面评估算法的性能。

1）以特征波形为单位，计算所有样本所有特征波形的阳性预测值PPV_Gross（以PPV_G标识）

$$PPV\_G = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FP_i)} * 100\% \quad \text{……（7）}$$

其中，i= 1~N，表示N个样本（数据），TPi表示第i个样本的特征波形的真阳性个数，FP_i表示第i个样本的特征波形的假阳性个数；

2）以样本（数据）为单位，计算所有样本灵敏度的加权平均值PPV_Average（以PPV_A标识）

$$PPV_i = \frac{TP_i}{TP_i + FP_i} * 100\% \quad \text{.....（8）}$$

$$PPV\_A = \frac{\sum_{i=1}^{N} PPV_i}{N} \quad \text{.....（9）}$$

其中，N表示样本（数据）总数，PPV_i表示第i个样本(数据)的阳性预测值。

7.1.1.4 F1 分数

F1分数是敏感度与阳性预测值的调和平均数，通常用于综合评估算法的敏感度和特异度，公式定义如下：

$$F1 = \frac{2 \times Sen \times PPV}{Sen + PPV} \quad \text{.....（10）}$$

7.1.1.5 马修斯相关系数

马修斯相关系数（Matthews Correlation Coefficient，MCC）本质是预测值与真实值的皮尔逊相关系数在二分类下的特例，衡量两者线性关联强度，尤其在数据分类不平衡时表现出色，同时考虑TP、TN、FP、FN，避免因类不平衡导致的评估偏差，比如QRS波分类任务中，正常QRS波占比远高于异常QRS波占比，此时可使用MCC与Sen、PPV以及F1分数等指标综合评价算法分类性能。

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} * 100\% \quad \text{.....（11）}$$

取值范围为[-1,1]。取值为1，表示算法性能完美（预测结果与金标准完全一致）；取值为0，表示算法可能随机猜测；取值为-1，表示完全反向预测（预测结果与金标准完全相反）。故其取值越接近于1，算法性能越好。

7.1.1.6 受试者工作特征曲线 ROC 及曲线下面积 AUC

以每一个检测结果作为可能的诊断界值，计算得到相应的真阳率(敏感度)和假阳率(1-特异度)，以假阳率为横坐标，以真阳率为纵坐标绘制而成的曲线。

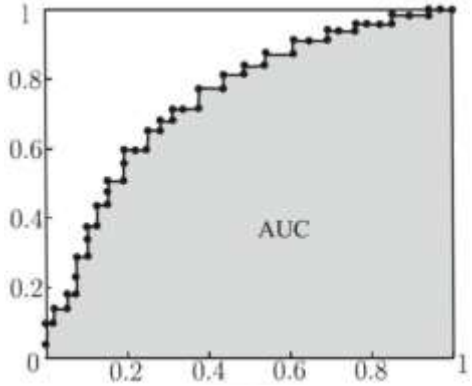

图1 ROC 曲线绘制示意图

计算曲线下面积，即为AUC。

7.1.2 心电波形特征参数自动测量

心电波形特征参数自动测量是心电图自动分析的重要领域之一，也是心电图临床诊断的重要依据。通常包含平均心率、P/QRS/T波的幅度、时限、间期等参数，如P波时限、QRS波时限、PR间期、QT间期。若算法支持自动测量，则应对各参数的测量误差进行统计，以评估算法的自动测量准确性。

针对每个测量参数，计算每例数据的测量误差，并统计数据集内的测量误差的平均值，即平均测量偏差，定义如下：

$$平均偏差\ \bar{e} = \frac{\sum_{i=1}^{N} e_i}{N} \quad \text{……（12）}$$

$$e_i = \hat{x}_i - x_i \quad \text{……（13）}$$

其中，N表示测试样本（数据）总数，e_i表示第i个样本的测量误差，$\hat{x}_i$表示第i个样本的算法测量值；x_i表示第i个样本的参考值（金标准）；

$$标准偏差 = \sqrt{\frac{\sum_{i=1}^{N}(e_i - \bar{e})^2}{N-1}} \quad \text{……（14）}$$

7.1.3 心电图自动诊断

7.1.3.1 心电图临床诊断结果分类概述

心电图的临床诊断，从结果上可按照不同分类标准及分类颗粒度分为不同类别，最简单可分为正常心电图和异常心电图。进一步的，异常心电图又包含诸多异常情况，如心房肥大、心室肥厚、心肌梗死、心律失常、起搏心电图等，每一类又可以根据位置不同、起源不同等进一步分类，常见的心电图诊断结果（术语）见附录B。因此，具有自动诊断功能的心电图辅助分析算法，通常是完成多分类任务，对应的，用于性能评测的数据库，就是多标签数据，所谓多标签，既指整个数据库中多个样本包含了多个标签，又指一个样本（一例数据）可同时包含多个诊断结果（标签）。

7.1.3.2 混淆矩阵

多分类任务，通常可以转化为多个单标签的二分类任务，并对每个标签的分类性能指标进行计算。在计算具体指标时，可根据算法诊断结果与金标准结果进行比对，完成混淆矩阵的统计。每个标签都可以形成二分类混淆矩阵：

表3 混淆矩阵示意表

二分类混淆矩阵		金标准	
		阳性	阴性
算法诊断结果	阳性	真阳性TP	假阳性FP
	阴性	假阴性FN	真阴性TN

7.1.3.3 单标签度量

针对每个可分析的输出结果（标签），进行单标签性能度量，采用以下指标。

$$Sen = \frac{TP}{TP + FN} * 100\% \quad \text{……（15）}$$

$$漏诊率 = 1 - \mathrm{Sen}$$

TP，真阳性样本的总数；FN， 代表假阴性样本的总数。

$$Spe = \frac{TN}{TN + FP} * 100\% \quad \text{……（16）}$$

$$误诊率 = 1 - \mathrm{Spe} \quad \text{……（17）}$$

TN，真阴性样本的总数；FP，代表假阳性样本的总数。

在心电图多标签自动诊断任务中，单标签的阴性样本往往显著多于阳性样本，因此，特异度常表现为较高水平，此时应同步使用误诊率及其他评价指标进行综合评估。

$$PPV = \frac{TP}{TP + FP} * 100\% \quad \text{……（18）}$$

TP，真阴性样本的总数；FP，代表假阳性样本的总数。

准确性是表示算法正确分类比率，关注点是正确分类的结果，其计算公式定义如下：

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} * 100\% \quad \text{……（19）}$$

在心电图多标签自动诊断任务中，单标签的阴性样本往往显著多于阳性样本，因此，准确性常表现为较高水平，可能弱化了算法错误分类（误诊及漏诊）的表现，应同时配合其他指标进行综合评价。

Matthews Correlation Coefficient（MCC）综合考虑TN、TP、FN、FP的结果，可以更好地综合评价敏感度和特异度,在阳性样本占比较低(如患病率低于5%的心脏疾病)等数据分类不平衡时表现出色，比ACC可以更好地体现算法平衡性。

指标计算公式参见公式：

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} * 100\% \quad \text{……（20）}$$

7.1.3.4 多标签（多分类）整体性能度量

心电图自动诊断需要同时识别多种异常（如“房颤+心肌梗死+室性早搏”），每个样本（数据）关联多个标签组合，因此需要综合考虑算法的多标签分类性能，此时应以单个样本（数据）为评估的颗粒度，评价算法的整体性能。

Kappa系数作为衡量多分类精度的指标可以比较好地反映多分类系统的综合性能,Kappa系数是基于混淆矩阵进行计算得到的。计算公式如下：

$$\mathrm{K} = \frac{P_0 - P_e}{1 - P_e} \quad \text{……（21）}$$

其中，Po是观测一致率，即每一类（标签）正确分类的样本数量之和除以总样本数。

假设每一类(标签)的真实样本个数分别为$a1$，$a2$，……，aC，而算法自动诊断（预测）出来的每一类(标签)的样本个数分别为$b1$，$b2$，……，bC，总样本个数为N，则有：

$$P_e = \frac{\sum_{i=1}^{M} (\hat{y}_i * y_i)}{N^2} \quad \text{……（22）}$$

其中，N为样本数量，M 为标签数量，$\hat{y}_i$表示第i个标签算法诊断出来的样本个数，y_i第i个标签真实的样本个数。

Kappa系数的范围为[-1,1],可以认为Kappa系数越接近1，分类精度越高；Kappa系数越接近-1，分类精度越低。

汉明损失(HL)衡量的是算法预测标签与真实标签之间的不符合率，即在所有样本的所有标签中，被错误预测的比例。汉明损失越小，表示算法的性能越好。

$$\mathrm{HL} = \frac{1}{N * M} \sum_{i=1}^{N} \sum_{j=1}^{M} I(\hat{y}_{ij} \neq y_{ij}) \quad \text{……（23）}$$

其中，N为样本数量，M 为标签数量，I为指示函数，当预测标签与真实标签不相等时为 1，否则为0。

宏观F1分数是对所有类别（标签）平等加权的F1分数算术平均值，用于平等观察所有算法诊断的心电图异常标签的综合性能。其计算过程如下：

假定算法可分析类别（标签）数量为M，首先针对每个类别（标签），逐类独立计算 F1分数 $F1_i$（基于该类别的Sen和PPV，或对应的TP、FP、FN）

$$F1_i = \frac{2 \times Sen_i \times PPV_i}{Sen_i + PPV_i} \quad \text{……（24）}$$

然后对所有类别（标签）的F1分数取均值，得到宏观F1分数：

$$\mathrm{Macro - F1} = \frac{\sum_{i=1}^{M} F1_i}{M} \quad \text{……（25）}$$

7.1.4 长时程心电事件检测

长时程心电图，通常用于检测一些持续性、非持续性以及阵发性心电事件，如心律失常、ST段改变等,此类心电图通常表现为既包含正常片段,又包含异常片段;同一类型的异常事件也可能有多个片段，且每一段的持续时长不等。针对此类事件，通常需要从两个方面进行性能测试：

a) 以每个片段为统计对象，统计片段的敏感度，阳性预测值；
b) 以时间的持续时间为统计对象，统计时间持续时间的敏感度和阳性预测值。

7.2 算法质量特性与测试方法

7.2.1 测试集\算法准确性不确定度

7.2.1.1 测试集不确定度

测试集不确定度满足：
a) 确定测试集的不确定度来源与分类：例如采集阶段设备误差、环境干扰误差，标注阶段主观误差、流程误差，扩增阶段偏移误差等；
b) 明确不确定度的量化评定方法：确定被测量与误差来源、量化各误差来源的标准不确定度、计算合成标准不确定度、计算扩展不确定度；
c) 不确定度的可接受阈值与控制措施：明确心电关键指标并标注不确定度，在各阶段制定措施控制不确定度。

7.2.1.2 算法准确性不确定度

算法准确性不确定度满足；
a) 确定算法准确度的不确定度来源与分类：例如测量误差、算法波动、统计误差、传递误差等；
b) 计算算法准确度不确定度：针对心电关键指标开展不确定度计算，例如准确率、召回率。

7.2.1.3 不确定度报告及追溯要求

不确定度报告及追溯要求包括：
a) 报告要求。标准测试集与算法准确性指标的不确定度需纳入测试报告，至少包含：不确定度评估方法、测试集不确定度明细、准确性指标不确定度明细、不确定度超标的原因分析与改进措施；
b) 追溯机制。应开展数据追溯、人员追溯、方法追溯，以保留不确定度计算的原始数据、明确不确定度评定的责任人、确保数据验证方法的科学性。

7.2.2 泛化能力

7.2.2.1 数据集构建原则

泛化能力测试应使用外部验证数据集。该数据集的构建应遵循以下原则：
a) 独立性：不得参与算法模型的任何开发阶段（包括训练、调参、特征选择及早期验证）。
b) 代表性：应尽可能覆盖真实世界应用中可能遇到的各种情况。
c) 规范性：数据应有经过临床金标准确认的、高质量的标注。
d) 溯源性：数据应保留充分的元数据信息，以便进行后续的亚组分析。

7.2.2.2 数据集多样性维度

外部验证数据集应从以下一个或多个维度体现多样性，并记录相关元数据：
a) 采集设备多样性：应包含来自不同生产厂商、不同型号、不同采样率、不同导联数量（如 12 导联、6 导联、3 导联、单导联）的心电设备采集的数据。
b) 人群多样性：应覆盖不同年龄、性别、种族、体重指数（BMI）、基础疾病（如高血压、糖尿病）的受试者。
c) 临床条件多样性：
   1) 应包含目标适应证相关的各种心律失常类型（如房颤、室性早搏、房室传导阻滞等）。
   2) 应包含可能存在的各种心电现象与干扰，如但不限于：基线漂移、工频干扰、肌电干扰、起搏器信号、低电压、高电压、心肌梗死波形等。
   3) 应包含合并多种心脏或非心脏疾病的复杂临床情况。
d) 采集环境多样性：应包含静息态、运动后、睡眠中、院内、家庭等多种场景下采集的数据。

7.2.2.3 亚组分型

应在整体测试的基础上，针对数据集多样性维度中定义的各个多样性维度进行亚组分析。例如：

a) 分别计算算法在来自 A、B、C 三家厂商设备数据上的性能指标。
b) 分别计算算法在不同年龄段（如<40 岁，40-60 岁，>60 岁）人群上的性能指标。
c) 分别计算算法对房颤、室早等不同心律失常类型的检测性能。

通过亚组分析，识别出算法性能可能显著下降的特定场景或人群。

7.2.3 鲁棒性

7.2.3.1 面向硬件变化的对抗测试

测试人员应考虑心电硬件设备、参数设置的多样性，收集或模拟生成更多的心电特征数据，作为对测试集的扩充，验证算法面对心电采集硬件设备的鲁棒性。

7.2.3.2 面向软件前处理的对抗测试

测试人员宜考虑软件前处理的多样性，收集或模拟生成更多的心电数据，作为测试集的扩充，验证算法面对软件前处理的鲁棒性。软件前处理的多样性包括:滤波处理、消除噪声、平滑预处理等。

7.2.3.3 面向欺骗攻击的对抗测试

欺骗攻击是一种加入人员难以觉察的扰动从而骗过模型的攻击手段，测试人员可使用白盒攻击(Projected Gradient Descent，PGD) 产生最大范数有限的扰动，并将扰动插入到原始心电数据中，然后用模型对这些添加扰动后的数据进行测试，从而验证模型是否能抵御恶意欺骗攻击。

测试人员宜根据产品的网络安全能力及风险分析文档，确定欺骗攻击的适用性和试验参数配置。施加扰动后的数据应通过标注人员的确认后用于测试。

7.2.3.4 压力测试

7.2.3.4.1 压力样本的定义

压力样本是指在心电分析算法模型的标定范围内，特征容量极大或者极小的样本。压力样本不应影响医生的正常判断。

7.2.3.4.2 压力样本的选取

压力样本的选取可遵循以下原则:

a) 受试者年龄偏大或者儿童（若适用）的心电图；
b) 特定疾病的心电图；
c) 有植入物(干扰项)的心电图，如起搏心电图；
d) 含有多种病变的复杂心电图。

7.2.4 重复性

7.2.4.1 测试环境

测试应在以下受控环境下进行：

a) 硬件环境固定：使用同一台计算设备（主机、处理器、内存等）。
b) 软件环境固定：使用完全相同的软件版本、依赖库、配置文件。
c) 输入数据固定：使用一组固定的、经过金标准确认的心电信号样本。

7.2.4.2 测试样本

应选取能代表算法预期用途的测试样本，包括：

a) 不同节律类型：应包含正常窦性心律和各种目标心律失常类型的样本。
b) 不同信号质量：应包含高质量信号和符合临床实际的、具有一定噪声的信号样本。

7.2.5 一致性

测试集要求包括：

a) 数据完整性：每例心电数据应包含完整的导联信息、采集参数、受检者基础信息、临床诊断结果，缺失关键信息的数据不得纳入测试。
b) 数据准确性：数据集应使用正式版本，临床数据需经多名专家审核，确保诊断数据的准确性。

7.2.6 效率测试

效率测试为在临床实际使用场景下，完成心电辅助分析任务的“时间性能”与“资源占用性能”的综合表现，核心体现为时间效率和资源效率。其中时间效率为软件从“完成心电数据导入”到“输出完整分析结果”（含分类结论、数值特征、异常提示）的总时间，不包含“数据传输时间；资源效率为软件分析过程中对硬件资源的消耗，包括CPU 使用率、内存占用量、显卡（GPU）使用率，需排除“系统后台程序”的资源消耗干扰。效率测试相关指标如下：

a) 时间效率：包含平均单次推理时间、吞吐量、P95/99 推理时间。
b) 资源效率：包含峰值下 CPU、内存、GPU 占用率，平均 CPU、内存、GPU 占用率。

7.2.7 错误分析

7.2.7.1 错误原因与归因

归因大类	具体子类	描述	示例
数据相关	信号质量差	因输入信号质量过低导致算法无法正常分析。	严重基线漂移、工频干扰、肌电噪声淹没特征波形。
	标注歧义/错误	训练或测试数据集的标注本身存在错误或高度主观性。	专家间对某一波形是否存在分歧；金标准标注错误。
	分布外数据	数据属于训练数据未充分覆盖的临床罕见情况或极端形态。	非常见的心律失常类型、罕见的心脏病并发症波形。
模型相关	特征学习不足	模型未能学会区分特定类型的特征。	无法有效区分房性早搏与室性早搏的形态。
	过度敏感/迟钝	模型对噪声或微小变化过度敏感，或对明显特征不敏感。	将T波误识别为P波（过度敏感）；遗漏低幅度的室性早搏（迟钝）。
	上下文理解错误	模型未能结合心电信号的上下文时序信息进行判断。	对心律的节律性判断错误，仅基于孤立波形进行分类。
后处理相关	规则逻辑缺陷	算法后处理阶段的决策规则或逻辑存在缺陷。	RR间期判断的阈值设置不合理，导致事件合并或分割错误。
	置信度阈值不当	输出概率的置信度阈值设置过高或过低，导致FP或FN增多。	房颤概率阈值设为90%，导致很多80%概率的房颤被漏判。

7.2.7.2 错误分析方法

错误分析方法包括：

a) 结果对比：将算法错误输出与参考标准（金标准）逐特征对比，明确偏差点（如 QRS 波群识别准确率、心律失常类型判断一致性）。
b) 场景对比：分析同一错误在不同场景下的发生频率，识别场景关联性错误。
c) 数据溯源：核查错误数据的质量，分析原始数据是否存在不可修复的质量问题；验证数据分布匹配度，对比错误数据与算法训练数据集的分布差异，判断数据场景训练覆盖不全导致错误。
d) 模型溯源：复盘算法模型的关键参数，通过控制变量法测试参数调整对错误率的影响。

e） 工程溯源：通过代码走查、单元测试，验证算法推理流程是否存在逻辑漏洞；通过在不同硬件配件下重复执行错误案例，判断是否因环境适配导致错误。

7.3 测试后封样

为保存测试状态证据，以便在监管审查、质量审计或对测试结果有争议的情况下提供原始依据，在全部指标测试完成后，需开展测试后封样，具体示意图如下：

图2 测试后封样示意图

其中封存数据应包含测试数据集全部元数据、测试设备、测试结果等信息，保证测试的可追溯性和可复现性。

附 录 A
（资料性）
心电图人工智能辅助判断测试数据集描述样例

A.1 数据采集适用范围

数据集适用于声称能对心电信号进行辅助分析的人工智能医疗器械软件产品。

A.2 数据采集

数据采集需考虑患者人群、采集场所、采集设备、数据格式、采集人员等方面的多样性，具有合规性证明，如伦理审批。表A.1给出了数据来源的多样性统计，可根据实际掌握的信息进一步细化。

表A.1 数据来源多样性统计

年龄	0~28天	Xx例
	29天~1岁	Xx例
	2岁~17岁	Xx例
	18~60岁	Xx例
	61~80岁	Xx例
	>80岁	Xx例
性别	男	Xx例
	女	Xx例
地域	华东地区	Xx例
	华南地区	Xx例
	华中地区	Xx例
	华北地区	Xx例
	西北地区	Xx例
	西南地区	Xx例
	东北地区	Xx例
心电采集设备型号	A公司xx型号	Xx例
	B公司xx型号	Xx例
导联数量	12导联	Xx例
	18导联	Xx例
采集场所	体检	Xx例
	门诊	Xx例
	住院	Xx例

表A.2 疾病分布多样性统计

心电图类别	亚组分类	样本量
正常心电图	\	Xx例
心房肥大	左房肥大	Xx例
	右房肥大	Xx例

	……	Xx例
心室肥厚	左心室肥厚	Xx例
	右心室肥厚	Xx例
	……	Xx例
心肌梗死	前壁心肌梗死	Xx例
	下壁心肌梗死	Xx例
	……	Xx例
ST-T改变	ST段改变	Xx例
	T波异常	Xx例
	……	Xx例
心律失常	室性早搏	Xx例
	房性早搏	Xx例
	……	Xx例
起搏心电图	心房起搏心律	Xx例
	心室起搏心律	Xx例
	……	Xx例

A.3 标注对象

特征波形标注：标记每个心搏的特征波形，如P\QRS\T波；

波形特征点及特征参数标注：定位P/QRS/T波起、止点，测量PR间期、QTc间期等参数。

诊断分类：基于心电图和/或结合临床信息（如超声、冠脉造影）标注心电图诊断结果（如“心肌梗死、心房颤动”），区分形态与节律异常。常见诊断结果参见附录B。

事件标注：标记心律失常、ST段抬高/压低等事件（如房颤发生片段及起、止时间），需精确至单个心搏。

标注对象的定义由心电图临床专家和工程技术专家组成的专家组给出，专家职称均为副高级以上，其中医疗系列专家从事临床工作的年限为10年以上，从事数据标注相关工作的年限为1年以上。

A.4 标注规则

组织3名心电图医生，根据制定好的标注标准，经培训后，使用软件背靠背标注心电图。

记录每名标注人员的标注结果。先采用少数服从多数法，即以不少于2名标注人员给出的该段信号的标注结果一致，则将其作为该段信号初始标注结果。

标注人员面对面复核信号初始标注结果，如对初始标注结果没有疑义，则初始标注结果即作为最终标注结果；如有疑义未达成一致结果，提请专家组仲裁(3位专家组成)，专家组结合初步标注结果，经讨论给出最终标注结果。

对于心电图的诊断结果，推荐使用附录B所述诊断术语进行标注。

A.5 标注人员

心电图医生从事临床工作的年限不低于1年，接受过标注规则培训。

仲裁专家组的职称不低于中级职称，从事临床工作的年限不低于8年，从事标注的年限不低于1年。人员的考核指标包括分类的准确率，要求不低于90%。

A.6 标注软件

标注软件推荐使用已上市产品完成；若使用自编软件，则需要提供软件功能说明及测试验证报告，确保标注过程及结果可靠。自编软件主要功能应至少包括心电数据的读取、显示、添加标注、标注审核与修改、保存标注结论。

A.7 标注环境

标注任务在具有资质医院或科研机构的医学人工智能实验室进行，使用医用显示器及办公电脑进行，无特殊环境要求。

附 录 B
（资料性）
心电图诊断术语

表B.1 首要诊断术语

A总述	1 标准心电图	2 其他正常心电图	3 异常心电图	4无法解释心电图
B技术条件	10 肢体导联反接	11 胸导联位置错误	12导联脱落	13 右心前区导联
	14 人工伪差	15 数据质量差	16 后壁导联	
C 窦性心律及心律失常	20 窦性心律	21 窦性心动过速	22窦性心律过缓	23 窦性心律失常
	24 一般窦房阻滞	25 二度窦房阻滞	26 窦性停搏	27不确定性室上性心律
D室上性心律失常	30 房性早搏	31房性早搏未下传	32折返性房性早搏	33不稳定性心房起搏
	34 心房异位节律	35多源性心房	36交界性早搏	37交界性逸搏
	38 交界性心律	39加速交界性心律	40室上性心律	41室上性
	42非窦性心动过缓			
E室上性心动过速	50 心房颤动	51心房扑动	52异位房性心动（单源性）	53异位房性心动过速(多源性)
	54 交界性心动过速	55室上性心动过速	56窄QRS波群、心动过速	
F室性心律失常	60 室性早搏	61融合波	62室性逸搏	63室性自主心律
	64加速性室性	65分支性节律	66并行收缩、自主心律	
G室性心动过速	70室性心动过速	71非持续性心动过速	72多源性室性心动过速	73尖端扭转型室性
	74心室颤动	75 分支性心动过速	76宽QRS波群心动过速	
H房室传导	80 PR间期过短	81房室传导比N :D	82 PR 间期延长	83二度房室传导阻滞,莫氏I型
	84 I度传导阻滞	85 2:1房室传导	86房室传导阻滞莫氏|型(不稳定传导)	87高度房室传导阻滞
	88三度房室传导阻滞	89 房室分离(完全性)		

I心室内及房内 传导	100室上节律异常传导	101左前分支阻滞	102左后分支阻滞	104左束支阻滞
	105不完全性右束支阻滞	106右束支阻滞	107室内差异性传导	108心室预激
	109右心房传导异常	110左心房传导异常	111 Ep sibn波	
J电轴与电压	120电轴右偏	121电轴左偏	122右上电轴	123不确定电轴
	124电交替	125低电压	128心前区R波异常增高P波电轴异常	131 P波电轴异常
K心腔肥厚及扩大	140左房肥大	141右房肥大	142左心室肥厚	143右心室肥厚
	144室间隔肥厚			
L ST段, T波,U波	145 ST段改变	146 ST-T改变	147 T波异常	148 QT间期延长
	149短QT间期	150 U 波高尖	151 插入性U波	152 T波U波融合
	153心室肥厚所致的ST-T改变	154 0 sborm波	155 过早复极	
M心肌梗死	160前壁心肌梗死	161下壁心肌梗死	162后壁心肌梗死	163侧壁心肌梗死
	165前间壁心肌梗死	166广泛前壁心肌梗死	173心肌梗死合并左束支阻滞	174右室心肌梗死
N起搏器	180心房起搏心律	181 心室起搏心律	182非右室心尖来源的心室起搏	183心房感知心室起搏波或节律
	184房室双腔起搏	185心房失夺获186心室失夺获	187心房感知不良	188心室感知不良
	189心房起搏不良	190心室起搏不良		

表B. 2 次要诊断术语

建议性术语	200急性心包炎	201急性肺动脉栓塞	202 Brugada综合征	203慢性肺血管疾病
	204中枢神经系统疾病	205洋地黄效应	206洋地黄中毒	207高钙血症
	208高钾血症	209甲亢性心肌病	210 低钙血症	211低钾血症或药物作用
	212低体温	213房间隔缺损	214心包积液	215窦房结功能障碍
考虑性术语	220急性心肌缺血	221房室结折返	222房室折返	222房室折返
	224高位心前区导联	225甲状腺机能减退	226缺血	227左心室室壁瘤
	228正常变异	229肺动脉疾病	230右位心	231右转位

参 考 文 献

[1] Yuanyuan Tian, Zhiyuan Li, et al. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. Cell Reports Medicine. https://doi.org/10.1016/j.xcrm.2024.101875

[2] 中国心电学会 中国心律学会 编译，心电图标准化和解析的建议与临床应用 国际指南2009 ，中国环境科学出版社

[3] GB/T9386—2008计算机软件测试文档编制规范

[4] 人工智能医疗器械注册审查指导原则 [国家药品监督管理局医疗器械技术审评中心（2022年第8号）]

[5] 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医疗器械技术审评中心（2019年第7号）]