ICS
CCS

T/CMA

中 国 计 量 协 会 团 体 标 准

T/CMA ZK 161—2024

测控装备智能化模型可信度计量技术

Reliability measurement technology for intelligent models of measurement and control equipment

2024－XX－XX 发布

2024－XX－XX 实施

中 国 计 量 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。

本文件由中国计量协会提出。

本文件由中国计量协会智库工作委员会归口。

本文件起草单位：

本文件主要起草人：

测控装备智能化模型可信度计量技术

1 范围

本文件规定了测控装备的智能模型可信度测评方法的一般要求、测评数据集选择标准、测评指标的选取、定义、计算方式、测评过程、测评方法、测评结果的不确定度等内容。

本文件适用于智能仪器仪表的智能模型精确度和可信度的测评，其他测控装备例如机器人、变频器、机床、变送器、控制系统等的智能模型精确度和可信度的测评可以参照本文件执行。

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 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

GB/T 29268.2-2012 信息技术 生物特征识别性能测评和报告 第2部分：技术与场景评价的测评方法

GB/T 41772-2022 信息技术 生物特征识别 人脸识别系统技术要求

GB/T 42888-2023 信息安全技术 机器学习算法安全评估标准

GB/T 42981-2023 信息技术 生物特征识别 人脸识别系统测评方法

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

JJF 1001-2011 通用计量术语及定义

JJF 1059-2012 测量不确定度评定与表示

3 术语和定义

GB/T 17 212、JJF1059-2012 界定的及以下术语和定义适用于本标准。

3.1

智能仪器仪表 Intelligent instrument

通过计算机、通信技术和人工智能等现代技术手段，利用先进的信息技术和传感技术，实现数据采集、处理和显示等功能，将仪器仪表传统的测量、控制、调节和溯源等功能进行升级，实现智能化、自动化和智能决策的新型设备。

3.2

测控装备 Measurement and control equipment

用于制造过程测量和控制的装备。

注 1:测控装备包括控制系统和仪器仪表。其中控制系统是用于数控机床、基础测控装备、流程工业装备及其他测控装备中实现控制功能的工业控制系统：仪器仪表是用于离散制造和流程工业装备中，连续测量温度、压力、流量、物位等变量，或者测量物体位置、倾斜、旋转等物性参数以及物质成分的仪器和仪

表。

注 2:测控装备包括装备硬件、嵌入式软件和智能化软件等。

3.3

精确度 Accuracy

仪表的示值与被测量(约定)真值的一致程度。

注：“精确度”是一个常用术语，它包含了系统误差、随机误差、回差、死区等影响。虽然把所有这些误差都组合在“精确度”这一标题下使用起来很方便，但它是一个定性的术语，其本身不赋数值。

[来源：GB/T 17212-1998 5]

3.4

可信度 Trust worthy

由以下参数定义：准确性、可靠性、弹性、客观性、有效性、可解释性、安全性、问责性和私密性。

3.5

测评集 Feature matching

输入一个系统或模型进行测评的数据集。

3.6

目标识别 Object recognition

识别图像中对象的类别。

注：目标识别也称作目标分类。

3.7

目标检测 Object detection

确认图像中是否存在指定类别的对象并确定其位置和大小。

3.8

几何变换 Geometrical transformation

一种改变像素空间位置的图像变换方法。

注：常见图像几何变换包括图像旋转、缩放、平移、镜像、错切、投影变换等。

3.9

图像变换 Image transformation

一种将原始图像空间数据(颜色、灰度)以某种形式映射到另一空间的方法。

注：包括图像空间域变换、频域、小波变换、DCT 变换等。

3.10

F1 测度 F1 score

精度和召回率的调和平均值。

注：调和平均值计算由 2 倍的精度乘以召回率除以精度和召回率的和得到。

3.11

精度　Precision

预测类别为正样本的集合中真实类别为正样本的比率。

注：精度一般每一类分别计算。

3.12

交叉验证　Cross validation

将数据集合划分成多组不相交的 n 份，每次使用 n—1 份作为训练集，另外一份作为验证集，重复 n 次。将 n 次结果平均(或者通过其他计算方式)得到最终结果的过程。

3.13

假负率　False negative rate

预测类别为负样本的集合中真实类别为正样本的比率。

3.14

可信　Trust

用户或其他利益相关方对产品或系统将按预期运行的置信程度。

[来源：ISO/IEC 25010:2011,4.1.3.2,有修改]

3.15

留一法　Leave one out

每次将一个(组)样本作为验证样本，其余 n-1 个(组)样本作为训练样本。重复实验 n 次，保证每个样本均曾作为验证。基于所有样本的预测结果计算最终评价结果的方法。

注：留一法是交叉验证的一个特例。

3.16

标准参考数据 Standard reference data

经过数据采集、数据清洗、整理筛选、专家评定的数据。

3.17

受试者工作特征曲线　Receiver operating characteristic curve

ROC 曲线

由不同设定条件下的真正率和假正率值画出的响应曲线。

3.18

ROC 曲线下面积　Area under ROC curve

ROC 曲线下的积分面积。

3.19

虚警率　False alarm

被错误预测为正样本的负样本占全部负样本的比率。

3.20

召回率　Recall

被正确预测的正样本占全部正样本的比率.

注 1:召回率和精度一般具有反比关系：一方升高时另一方趋向于降低。

注 2:也称为真阳性率。

[来源：ISO 25964-2:2013,3.65,有修改]

3.21

真负率　True negative rate

被正确预测的负样本占全部负样本的比率。

3.22

准确率　Accuracy rate

预测类别和真实类别相同的的样本数占全部样本数的比率。

3.23

测量不确定度 Measurement uncertainty,uncertainty of measurement

根据所用到的信息，表征赋予被测量值分散性的非负参数。

注：简称不确定度。

【来源：JJF 1001,5.18】

3.24

标准测量不确定度 standard measurement uncertainty,standard uncertainty of measurement

以标准偏差表示的测量不确定度。

注：简称标准不确定度。

【来源：JJF 1001,5.19】

4　一般要求

4.1　概述

智能测控装备主要包括数控机床、金属切割和焊接设备、仪器仪表、工控自动化等产品，智能测控装备工作过程中运用互联网、软件工程技术、自动化控制等新兴技术，相比于传统测控装备，其呈现“自动化”“智能化”特征，具有感知、分析、推理、决策和控制功能。

测控装备的智能水平直接影响其应用效果。评价测控装备的智能水平需首先评价其智能模型。目前大多测控装备面临智能化水平参差不齐，亟须通过对其智能模型进行测评实现对智能化水平的评价，智能模型的各类性能中，精确度与可信度是关键指标。智能模型的精确度、可信度决定了测控装备智能水平的基本水准。因此需首先规范测控装备智能模型的测评方法。基于计量数字化技术和数字计量技术，本标准建立了测控装备智能模型精确度测评方法（可信度适用），包括相关术语、规范化表达、测评流程、测评方

法。通过制定测控装备智能模型测评标准，形成规范化、受认可的计量标准文件，从而规范测评流程、方法、测评指标与测评技术。

4.2 测评环境

除特殊规定外，基本测评环境要求如下：

a) 环境温度：15℃～35 ℃；

b) 相对湿度：25%～75%；

c) 环境噪声：小于60 dB；

d) 硬件平台：测试系统的性能应满足被测系统运行基本要求，如CPU主频大于2.0 GHz、内存大于4GB或等同性能的硬件平台；

e) 操作系统：应为主流操作系统；

f) 进行测试时，应提供测试接口函数，测评接口函数按照附录A；

g) 其他测试工具等。

4.3 测评数据集

4.3.1 概述

测评数据集用于测控装备智能化模型精确度与可信度的测评，通常选择标准数据集或参考数据集为测评数据集，测评数据集应符合计量测试规范并带有计量属性，针对不同测控装备的不同种类智能模型应选择符合该智能模型测评领域的对应数据集，智能模型对应数据集举例见表1。

测评数据集依据数据集来源分为生成数据集、公开数据集、采集数据集。

表 1 不同种类智能模型对应数据集举例

序号	测控装备名称	智能模型	对应测评数据集
1	机器人	目标检测模型	目标检测模型测评数据集
2	仪器仪表	目标识别模型	目标识别模型测评数据集
3	变频器	智能控制模型	智能控制模型测评数据集
4	数控机床	故障诊断模型	故障诊断模型测评数据集
5	PLC	定位控制模型	定位控制模型测评数据集
6	阀门	自适应控制模型	自适应控制模型测评数据集

4.3.2 生成数据集

可通过AI手段生成测评数据集。例如通过深度学习或人工智能生成对应数据集。

4.3.3 公开数据集

可选择MINST数据集、CIFAR-10、ImageNet数据集等经典公开、开源数据集。

4.3.4 采集数据集

通过数据收集获取真实场景的真实数据集。

4.3.5 测评数据集要求

测评数据集应符合以下要求：

a）数据集为参考数据集、标准数据集；

b)数据集中数据满足标准数据、参考数据、计量数据之一；

注：参考数据：依照数据质量评估流程得到的数据；标准数据：来自国家标准、行业标准、地方标准的数据；计量数据：经过计量器具或标准、软件认定、计量过的数据，其中涉及计量方法与计量数据均带有计量属性。

c)数据集覆盖率满足要求；

d)数据集标注清晰，检查标注质量合格；

e)数据集格式、数量满足要求；

f)数据集质量满足要求，如图片数据无遮挡、图片清晰度满足要求、声音数据无噪声；

g)数据集存储设备应由专人负责保管。

5.测评指标

5.1 测评指标选择

以下七个指标作为测控装备智能模型精确度与可信度测评指标：accuracy（准确率）、F1-measure（F1分数）、Precision（精确率）、IoU（重叠度）、ROC（受试者工作特征曲线）、AUC（受试者工作特征曲线包络面积）、recall（召回率）。

a）F1分数（F1 score）。又称平衡F分数（balanced F Score），它被定义为精确率和召回率的调和平均数：

$$\mathrm{F1}=\frac{2}{\frac{1}{\mathrm{P}}+\frac{1}{\mathrm{R}}}$$

F1分数通过赋予较小值更大的权重，避免较小值和较大值对结果产生较大影响，综合考虑了模型的预测准确性和召回能力，因此相较于单一的查全率和查准率具备更好的评估效果。

b)准确率（Accuracy）。即预测正确的结果占总样本的百分比：

$$\mathrm{Acc}=\frac{TP+TN}{TP+TN+FP+FN}$$

c)精准率（Precision，查准率）。表示为预测正确的正例数据占预测为正例数据的比例，用于衡量模型的预测准确性：

$$\mathrm{P}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

相比于准确率，精准率只针对预测正确的正样本而不是所有预测正确的样本，其结果表示预测结果为正样本中真正为正样本的比例，也成为为查准率，反映了模型对正类别的识别能力。

d)召回率（Recall，查全率，敏感度）。表示为预测为正例的数据占实际为正例数据的比例：

$$\mathrm{R}=\frac{TP}{TP+FN}$$

e)ROC曲线（Receiver Operating Characteristic Curve，受试者工作特征曲线）,ROC曲线的横轴是“假阳率”（False Positive Rate，FPR），纵轴是“真阳率”（True Positive Rate，TPR）；

ROC曲线横轴代表被预测为正样本的负样本占全部负样本的比率(假正率),纵轴代表被预测为正样本的正样本占全部正样本的比率(召回率)。召回率越高，假正率越低，则模型性能越好。

f)AUC则表示ROC曲线下的面积，通常大于 0.5 小于 1。AUC值越大的分类算法，其性能越好：

$$AUC = (1 + Sensitivity - Specificity) / 2$$

g)IoU(Intersection over Union)：用于评估注释、分割和对象检测算法的准确性。它量化数据集中的预测边界框或分段区域与地面实况边界框或注释区域之间的重叠。IOU提供了预测对象与实际对象注释的匹配程度的衡量标准，从而可以评估模型准确性并微调算法以改进结果。

$$IOU = \text{交集面积} / \text{并集面积}$$

5.2 指标评估方法

对5.1中测评指标进行权重赋值，通过对各指标值加权求和得到测评的最终结论。权重赋值方法可参考以下内容：

a)客观赋值法

使用已有权重赋值算法，包括但不限于熵值法、模糊综合评价法、灰色关联法等。

b)主观赋值法

根据具体应用场景中影响因素的重要程度人为进行权重赋值，包括但不限于专家咨询法、评级打分法、AHP法等。

c)混合赋权法

主观赋权与客观赋权方法结合，既保证赋权客观可信，又最小化赋权的不确定度，提升智能模型可信度综合评估的准确性。

6.测评方法

测控装备智能化模型测评方法可采用以下方法：

a)留出法：是最简单、最基础的一种数据集划分方法。它将原始数据集划分成两个互斥的集合，一个作为训练集，另一个作为测试集。具体操作时根据实际需求把原始数据按比例分成两部分：一部分用来训练模型，另一部分用于验证智能模型。

b)交叉验证法：将全部数据集进行随机、均等切割，并多次使用不同子集做为模型验证数据。在此基础上计算平均值以减少因为子集选择所带来的误差或偏差。

c)自助法：假设给定的全部数据集包含d个样本。该数据集有放回地抽样d次，产生d个样本的训练集。这样原数据样本中的某些样本很可能在该样本集中出现多次。没有进入该训练集的样本最终形成检验集（测试集）进行模型评估。

d)测试时增强（TTA：Test-time augmentation methods）：通过对测试数据集样本进行多次随机变换或扰动，产生多个增强的样本，并使用这些样本进行预测的多数投票或平均来得出最终预测结果。

在实际测评任务中，根据具体情况来选择合适的测评数据集与测评方法。

7.测评过程

测评过程见图 1：测评流程图。

图1 测评流程图

7.1 明确测评任务

首先需明确测评对象和应用场景，了解用户测评需求和测评对象，例如测控装备智能化模型的具体类型：算法类型、运行环境、开发框架、算法语言和结构等，基于测评对象和测评需求的分析，明确测评任务，依据任务选取测评方法。

7.2 选取测评指标

测评指标见5.1。

7.3 选取测评数据集

a) 测试数据集需满足4.3.5中测评数据集的要求；

b) 测试数据集需与测评对象的智能模型相匹配；

c) 测试数据集应具备多样性；

d) 测试数据集中不包含模型训练数据集中的数据；

e) 测试数据集中需包含一定数量的干扰数据和对抗数据、例如OOD数据（目标域之外数据）、加噪声数据、旋转或变换后的图片等。

7.4 准备测评资源与环境条件

为测评任务的开展和实施做准备，配置相关资源和环境：准备测评数据、测评指标、测评方法、测评模型、软硬件运行环境等。

7.5 实施测评

在配置好的测评资源与环境下开始测评，将准备好的测评数据集输入待测智能模型，监测测评过程并获取过程数据和结果数据。

7.6 数据采集

收集并保存实施测评过程中产生的数据。

7.7 计算测评指标

根据指标计算公式，代入测评数据集与待测模型输出结果计算测评指标数值。

7.8 结果综合评估

结合5.2中指标评估方法，得出最终的测评结果。

7.9 测评报告

综合测评流程及其结果，出具相应的测评报告。

7.10 测评结果

测评原始记录参考格式按照附录B。

附 录 A
（规范性）
测评示例

A.1 文件格式

接口函数采用C/python语言开发时，支持多线程，可编译为32位或64位版本。

A.2 接口函数

测评接口函数见表A.1。

表 A.1

序号	函数名称	说明
1	Evaluation_inputdata	数据集引入接口
2	Evaluation_inputindex	测评指标选择并计算接口
3	Evaluation_inputmethod	测评方法选取接口
4	Evaluation_inputmodel	测评对象引入接口
5	Evaluation_outputresult	结果输出并显示数值
6	Evaluation_storageresult	结果存储
7	Evaluation_Visualizationresult	结果可视化

A.3 接口函数说明

各个测试接口的定义：

（1）数据集引入接口：

编号：1
接口名称：Evaluation_inputdata
接口功能：用于实现测评数据集的接入
接口输入参数符号：inputdata
接口输入参数含义：测评数据集
接口输入参数数据类型：图像数据或其他数据类型
通信协议：TCP/IP协议，HTTP协议

（2）测评指标选择并计算接口：

编号：2

接口名称：Evaluation_inputindex
接口功能：选择计算可信度测评指标并实施计算过程。
接口输入参数符号：cal_index
接口输入参数含义：计算测评指标
接口输入参数数据类型：无
通信协议：TCP/IP协议，HTTP协议

（3）测评方法选取接口：

编号：3
接口名称：Evaluation_inputmethod
接口功能：用于引入选择的测评方法，此处选择某个测评方法进行测评。
接口输入参数符号：cal_method
接口输入参数含义：选择测评方法
接口输入参数数据类型：无
通信协议：TCP/IP协议，HTTP协议

（4）测评对象引入接口：

编号：4
接口名称：Evaluation_inputmodel
接口功能：用于接入测评对象：测控装备智能模型
接口输入参数符号：input_model
接口输入参数含义：引入测评模型
接口输入参数数据类型：无
通信协议：TCP/IP协议，HTTP协议

（5）结果输出并显示数值：

编号：5
接口名称：Evaluation_outputresult
接口功能：用于输出计算后得到的指标结果并显示。
接口输出参数符号：out_result
接口输出参数含义：输出并显示指标。
接口输出参数数据类型：字符串
通信协议：TCP/IP协议，HTTP协议

（6）结果存储：

编号：6
接口名称：Evaluation_storageresult
接口功能：用于存储计算后得到的测评指标等数据
接口输出参数符号：store_result
接口输出参数含义：存储结果
接口输出参数数据类型：字符串
通信协议：TCP/IP协议，HTTP协议

（7）结果可视化：

编号：7
接口名称：Evaluation_Visualizationresult
接口功能：用于将测评指标的计算结果输入可视化工具后进行可视化展示
接口输出参数符号：vision_result
接口输出参数含义：测评结果
接口输出参数数据类型：图像数据
通信协议：TCP/IP协议，HTTP协议

A.4 不确定度评定

A.4.1 环境条件：

温度：21.0℃，相对湿度：42%。

A.4.2 测评数据集：

温湿度表盘数据集。

A.4.3 被测对象：

智能温湿度仪表表盘读数识别模型。

A.4.4 测量方法：

将温湿度表盘数据集作为测评数据集输入智能温湿度表表盘读数识别模型，测评方法选择测试时增强法（TTA），计算测评指标，收集指标计算结果并通过指标评估方法分析，进行精确度与可信度评估。

A.4.5 不确定度来源：

a)测评方法引入的不确定度分量V_1
b)智能模型带有不确定度分量V_2
c)测评数据集带有不确定度分量V_3
d)指标评估方法引入的不确定度分量V_4

最终结果的不确定度汇总为u_i 。

A.4.6 标准不确定度评定

智能温湿度表表盘读数识别模型本身带有不确定属性，且温湿度表盘数据集带有不确定度，测评方法的选择会在测评过程引入不确定度，指标评估方法不同也会引入不确定度。

A.4.7 合成标准不确定度的计算

不确定度分量的汇总见表A.1。

表A.1 不确定度分量汇总表

u_i	u_i的来源	u_i的值	C_i	$\|C_i\|\times u_i$	自由度 υ_i

$u(V_1)$	测评方法引入的 不确定度分量	0.16	1	0.16	∞
$u(V_2)$	智能模型带有 不确定度分量	0.24	1	0.24	∞
$u(V_3)$	测评数据集带有 不确定度分量	0.12	1	0.12	∞
$u(V_4)$	指标评估方法引入的 不确定度分量	0.08	1	0.08	∞

合成标准不确定度为：

$$u_c=\sqrt{u^2(v_1)+u^2(v_2)+u^2(v_3)+u^2(v_4)}=\sqrt{0.16^2+0.24^2+0.12^2+0.08^2}=0.32$$

A.4.8 扩展不确定度

A.4.8.1 计算扩展不确定度

取置信概率p=95%，$v_{\text{eff}}\to\infty$，查t分布表得k= 2

扩展不确定度U= ku_c = 2×0.32 = 0.64

相对扩展不确定度为：$U_{rel}=\dfrac{U}{100V}\times100\%=0.64\%$

A.4.8.2 其它测控装备智能模型的不确定度

按上述方法可以求得其它测控装备智能模型精确度与可信度测评结果的不确定度。

附 录 B
（规范性）
测评原始记录参考格式

测评原始记录参考格式见表B.1。

表B.1 测评原始记录参考格式

测控装备智能模型	序号	测评环境条件	测评数据集	测评指标结果	测评指标评估分析	可信度结果	扩展不确定度
目标识别模型	1						
	2						
	3						
	4						

参 考 文 献

[1] GB/T 2900.13 电工术语 可信性与服务质量RENRENDOC.COMGB/T 3358(所有部分) 统计学词汇及符号

[2] GB/T 5080.4 设备可靠性试验 可靠性测定试验的点估计和区间估计方法(指数分布)

[3] GB/T 13983-1992 仪器仪表基本术语

[4] GB/T 17212-1998 工业过程测量和控制 术语和定义

[5] GB/T 17614.1 工业过程控制系统用变送器 第1部分：性能评定方法

[6] GB/T 18271.1 过程测量和控制装置 通用性能评定方法和程序 第1部分：总则

[7] GB/T 33767.5-2018 信息技术 生物特征样本质量 第5部分：人脸图像数据

[8] GB/T 41864-2022 信息技术 计算机视觉 术语

[9] JB/T 6214—1992 仪器仪表可靠性验证试验及测定试验(指数分布)导则

[10] JB/T 50123 仪器仪表现场工作可靠性、有效性、维修性数据收集指南