ICS 35.240
L70/84

团 体 标 准

T/CESA 1036—2019

信息技术 人工智能 机器学习模型及系统的质量要素和测试方法

Information technology - Artificial intelligence -Quality elements and testing methods of machine learning model and system

2019－04－01 发布 2019－04－01 实施

中国电子工业标准化技术协会 发布

目　　次

前　　言 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　缩略语 4
5　概述 5
6　质量要素 5
6.1　功能性 5
6.2　可靠性 6
6.3　效率 6
6.4　维护性 7
7　测试方法 7
7.1　标准数据集 7
7.2　功能性 10
7.3　可靠性 13
7.4　效率 16
7.5　维护性 16
附录 A　（资料性附录） 18
A.1 机器学习模型及系统测试实例 18

前　言

本标准按照GB/T　1.1－2009《标准化工作导则　第1部分：标准的结构和编写》给出的规则起草。

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。

本标准由中国电子技术标准化研究院提出并归口。

本标准起草单位：北京航空航天大学、中国电子技术标准化研究院、无锡市信息化协会、第四范式（北京）技术有限公司、中国医学科学院生物医学工程研究所、北京深醒科技有限公司、中国航空综合技术研究所、北京国华恒源科技开发有限公司、上海腾梭科技有限公司、浪潮软件集团有限公司、重庆邮电大学、电子科技大学、北京邮电大学、重庆中科云从科技有限公司、深圳云天励飞技术有限公司、南京行者易智能交通科技有限公司、海尔优家智能科技（北京）有限公司、南京中兴新软件有限责任公司、广州广电运通金融电子股份有限公司、中国电子科技集团公司第十四研究所、北京航天自动控制研究所、合肥中科类脑智能技术有限公司、北京京东尚科信息技术有限公司、华夏芯(北京)通用处理器技术有限公司、威麟信息技术开发（上海）有限公司、玉养信息科技（上海）有限公司、中国电信集团有限公司、上海交通大学苏州人工智能研究院、苏州思必驰信息科技有限公司、浙江大华技术股份有限公司。

本标准主要起草人：刘祥龙、吴文峻、代红、董建、张群、王燕妮、马珊珊、汪小娟、王挺、郭夏玮、王嘉磊、肖羽、蒲江波、徐圣普、肖鑫、田永会、王洁萍、王广、李静、张涛、胡亮、罗小勇、许浩、吴艳、王功明、黄先芝、黄庆卿、张焱、罗光春、田玲、张栗粽、王枞、闫敏、李军、胡文泽、王孝宇、程冰、孙雨新、林震亚、杜新凯、林冠辰、陈良旭、杨祎、孙晶明、王丽娜、徐颂、褚海涛、郑歆慰、刘海峰、李文慧、张振庭、刘军、翁家良、朱兆颖、杨震、李洁、俞凯、钱彦旻、程淼。

信息技术　人工智能　机器学习模型及系统的质量要素和测试方法

1　范围

本标准规定了机器学习模型及系统的质量要素，提供了机器学习模型及系统的质量测试指标体系以及相应的测试方法。

本标准适用于机器学习模型及系统的设计、研发及质量评价，用户可根据具体的机器学习模型选择合适的质量测试指标。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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　11457　信息技术　软件工程　术语

GB/T　16260.1-2006　软件工程　产品质量　第1部分：质量模型

GB/T　29831.1　系统与软件功能性　第1部分：指标体系

GB/T　29831.2-2013　系统与软件功能性　第2部分：测试方法

GB/T　29831.3　系统与软件功能性　第3部分：测试方法

GB/T　29832.1　系统与软件可靠性　第1部分：指标体系

GB/T　29832.2　系统与软件可靠性　第2部分：测试方法

GB/T　29832.3　系统与软件可靠性　第3部分：测试方法

GB/T　29833.1　系统与软件可移植性　第1部分：指标体系

GB/T　29833.2　系统与软件可移植性　第2部分：测试方法

GB/T　29833.3　系统与软件可移植性　第3部分：测试方法

GB/T　29834.1　系统与软件维护性　第1部分：指标体系

GB/T　29834.2　系统与软件维护性　第2部分：测试方法

GB/T　29834.3　系统与软件维护性　第3部分：测试方法

GB/T　29835.1　系统与软件效率　第1部分：指标体系

GB/T　29835.2　系统与软件效率　第2部分：测试方法

GB/T　29835.3　系统与软件效率　第3部分：测试方法

GB/T　29836.1　系统与软件易用性　第1部分：指标体系

GB/T　29836.2　系统与软件易用性　第2部分：测试方法

GB/T　29836.3　系统与软件易用性　第3部分：测试方法

GB/T　32904-2016　软件质量量化评价规范

3　术语和定义

GB/T　11457和GB/T　16260.1-2006界定的以及下列术语和定义适用于本文件。

3.1

机器学习模型 machine learning model

采用机器学习方法建立的输入与目标输出联系的计算模型。主要包含算法、超参数、参数、模型输入规范、模型输出规范五大要素。

3.2

模型输入规范 model input specification

规定模型输入的数据类型、数据格式、数据精度等。

3.3

模型输出规范 model output specification

规定模型输出的数据类型、数据格式、数据精度等。

3.4

机器学习模型系统 machine learning model system

机器学习模型的软硬件实现系统，可以保障模型在接受合适的数据输入后，可以正常运行，在规定时间内返回约定格式的输出数据。

注：本标准中，机器学习模型及系统是机器学习模型及机器学习模型系统的简称。

3.5

标准数据集 standard data set

符合一定规范要求的数据集，主要用于训练、验证和测试特定机器学习模型。

3.6

训练 training

对于给定的数据集，生成和优化机器学习模型参数设置的过程。

3.7

训练集 training set

用来训练的数据集。

3.8

测试 testing

对于给定的数据集，采用训练机器学习模型进行预测，由此评估训练模型性能的过程。

3.9

测试集 testing set

用来测试的数据集。

3.10

验证 validation

对于给定的数据，采用被验证模型进行预测，由此进行选择并优化训练模型结构和超参数的过程。

3.11

验证集 validation set

对于用来验证的数据集，称为验证集。

3.12

收敛 convergence

对于给定的数据集，机器学习模型训练达到局部最优或者全局最优的状态。

3.13

准确率 accuracy

对于给定的数据集，预测正确的样本占总样本的比率。

3.14

错误率 error rate

对于给定的数据集，预测错误的样本占总样本的比率。

3.15

精确率 precision

对于给定的数据集，预测为真正例的样本占预测为正例的样本的比率。

3.16

召回率 recall

对于给定的数据集，预测为真正例的样本占所有实际为正例样本的比率。

3.17

F1值 F1-score

精确率和召回率的调和平均。

3.18

受试者操作特性曲线 receiver operating characteristic

以假正例率为横坐标，真正例率（召回率）为纵坐标所组成的坐标图，和被试样本在特定刺激条件下由于采用不同的判断标准得出的不同结果画出的曲线。

3.19

平均绝对误差 mean absolute error

所有单个观测值与算术平均值的偏差的绝对值的平均。

3.20

均方误差 mean-square error

观测值与真值偏差平方和的平均值。

3.21

均方根误差 root mean square error

观测值与真值偏差的平方和与观测次数比值的平方根。

3.22

离群点 outlier

一个或一组明显不同于其他数据的数据点。

3.23

噪声数据 noisy data

错误或包含随机误差的数据。

3.24

拟合优度 goodness of fit

模型的预期值和现实所得的实际值的差距。

3.25

运行时间 running time

完成规定任务所需要的时间。

3.26

纯度 purity

正确聚类的样本数占总样本数的比例。

3.27

危险 hazard

机器学习算法发生失效，从而导致机器学习系统出现的一个非预期或有害的行为，或者提交给其他与机器学习系统相关联的系统发生错误。

3.28

危险严重性 hazard severity

某种危险可能引起的事故后果的严重程度。

4 缩略语

下列缩略语适用于本文件。

CPU：中央处理器（Central Processing Unit）

GPU：图形处理器（Graphic Processing Unit）

ROC：受试者操作特性曲线（Receiver Operating Characteristic）

5 概述

本标准依据GB/T 29831.1，GB/T 29832.1，GB/T 29833.1，GB/T 29834.1，GB/T 29835.1，GB/T 29836.1，GB/T 32904-2016综合提出机器学习模型及系统质量指标体系。本标准给出机器学习模型及系统的质量要素，其质量指标体系划分为功能性、可靠性、效率和维护性等4个主要特性及其子特性。

机器学习模型从已知数据中习得，并对未知数据进行预测，数据的质量制约着机器学习模型及系统质量。本标准给出了标准数据集的要求和测试方法，以标准数据集为基准在机器学习模型训练和测试过程中进行质量测试。机器学习模型及系统的质量指标体系见图1。

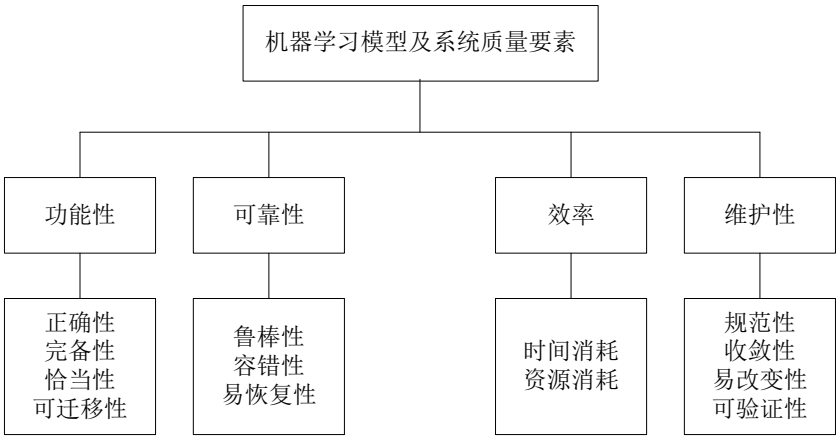

图1 机器学习模型及系统的质量要素

6 质量要素

6.1 功能性

6.1.1 概述

功能性用于定义和评价机器学习模型及系统满足用户对功能需求的能力。根据GB/T 16260.1-2006，GB/T 29831.1，GB/T 29831.2-2013，GB/T 29831.3，GB/T 32904-2016和机器学习模型的特点，功能性被分成正确性、完备性、恰当性，可迁移性等子特性。

6.1.2 正确性

正确性表明机器学习模型及系统对指定的任务和用户目标运行过程及产生结果的正确程度。它包含数据精度的满足性、模型设计的正确性、代码实现的正确性、计算结果的正确性等测试元。

6.1.3 完备性

完备性表明机器学习模型及系统对指定的任务和用户目标的覆盖程度。它主要包含功能实现与需求覆盖比、实现功能正交性等测试元。

6.1.4 恰当性

恰当性用于定义和评价机器学习模型及系统选择不同部件实现需求的合理性。它主要包含数据处理恰当性、模型设计恰当性、优化算法恰当性、模型实现恰当性、参数设置恰当性、训练操作恰当性等测试元。

6.1.5 可迁移性

可迁移性用于定义和评价机器学习模型及系统的迁移能力。它包含不同规模数据的可扩展性、同领域的可迁移性和不同领域可迁移性等测试元。

6.2 可靠性

6.2.1 概述

可靠性用于定义和评价机器学习模型及系统在规定条件下和规定时间内，完成规定功能的能力。根据GB/T 29832.1，GB/T 29832.2，GB/T 29832.3，GB/T 29833.1，GB/T 29833.2，GB/T 29833.3，GB/T 32904-2016和机器学习模型的特点，可靠性分为鲁棒性、容错性、易恢复性等子特性。

6.2.2 鲁棒性

鲁棒性用于定义和评价机器学习模型及系统避免异常和极端情况等危害导致失效的能力。它包含危害检出性、抗攻击性、抗干扰性等测试元。

6.2.3 容错性

容错性用于定义和评价机器学习模型及系统在发生故障时，维护用户期望的性能水平的能力。它包含失效的避免性、误操作的抵御性、误操作的危害性等测试元。

6.2.4 易恢复性

易恢复性用于定义和评价机器学习模型及系统发生失效时，在满足一定要求的时间内重新达到规定的功能，并恢复受影响的数据的能力。它包括平均恢复时间、易重新启动性、易复原性、复原的有效性等测试元。

6.3 效率

6.3.1 概述

效率用于定义和评价相对于所使用的资源，机器学习模型及系统完成工作的能力。其资源包括系统的软件和硬件配置、消耗资源（如CPU、GPU、内存、存储、能量等）和花费的时间。根据GB/T 29835.1，GB/T 29835.2，GB/T 29835.3，GB/T 32904-2016和机器学习模型的特点，效率分为时间消耗、资源消耗等特性。

6.3.2 时间消耗

时间消耗特性用于定义和评价在相同软件和硬件环境下，机器学习模型及系统训练和测试的时间消耗。在训练阶段，它包含模型收敛时间、模型训练单轮时间等测试元。在测试阶段，它包含模型执行一轮的时间等测试元。

6.3.3 资源消耗

资源消耗特性用于定义和评价机器学习模型及系统训练及运行时对硬件资源的消耗。它包含算法本身所需要的存储（硬盘、内存、显存等）占用、带宽（硬盘吞吐、网络流量等）占用及计算资源（CPU、GPU等）占用等测试元。

6.4 维护性

6.4.1 概述

维护性用于定义和评价机器学习模型及系统易于维护的程度。根据GB/T 29834.1,GB/T 29834.2，GB/T 29834.3，GB/T 29836.1，GB/T 29836.2，GB/T 29836.3，GB/T 32904-2016和机器学习模型的特点，维护性分为规范性、收敛性、易改变性和可验证性等子特性。

6.4.2 规范性

规范性用于定义和评价机器学习模型及系统的训练、运行及维护等阶段是否满足模型的规范标准。它包含模型设计的规范性、模型训练的规范性、模型测试的规范性、系统代码的易读性、系统版本兼容性等测试元。

6.4.3 收敛性

收敛性用于定义和评价机器学习模型的训练过程能否快速收敛达到预期性能。它包含模型收敛的稳定性、收敛时间以及收敛值等测试元。

6.4.4 易改变性

易改变性用于定义和评价维护者或用户对机器学习模型及系统进行修改、验证的难易程度。这些修改包括对机器学习算法代码的修改和对设计文档的修改。易改变性包含变更说明文档的完整性、模块间的耦合性、变更模块的可验证性等测试元。

6.4.5 可验证性

可验证性用于定义和评价机器学习模型及系统的计算过程及计算结果是否易于理解和验证。它包含模型计算过程的可验证性、模型计算结果的可解释性、系统功能的可验证性等测试元。

7 测试方法

7.1 标准数据集

7.1.1 划分特性

标准数据集应包含训练集（包含训练和验证两部分）和测试集。训练集用于机器学习模型的训练和生成，测试集用于测试机器学习模型的性能。标准数据集划分特性包含训练集和测试集的互斥性、样本分布一致性等子特性。

标准数据集的划分可采用“留出法（Hold-out）”、“交叉验证法（Cross Validation）”和“自助法（Bootstrapping）”等方法。训练集和测试集的互斥性、样本分布一致性等测试元的测试方法见表1。

表1 标准数据集的划分及测试方法

方法名称	方法介绍	测试公式	划分概率
留出法 (Hold-out)	将标准数据集划分为互斥的集合，训练/验证/测试集的划分要尽可能保持数据分布的一致性。	$D = S \cup T$，$S \cap T = \phi$ 式中： D——初始数据集； S——训练集； T——验证/测试集 （若测试集T样本量为1，则为留一法）。	一般情况下训练集、验证集、测试集比例为6:2:2；数据量（百万以上）比例可调整为98:1:1。
交叉验证法 (Cross Validation)	将标准数据集划分为k个大小相似的互斥子集。	$D = D_1 \cup D_2 \cup \Lambda \cup D_k$ $D_i \cap D_j = \phi(i \neq j)$ 式中： D——初始数据集； D_i——第i个子集。	k通常取10，表示10折交叉验证。
自助法 （Bootstrapping）	初始数据集大小为n，有放回抽样n次作为训练集。 n次都未抽中样本作为测试集。	$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = \frac{1}{e} \approx 0.368$ 式中： n——样本数量。 该公式表示未抽中的样本概率。	初始数据集D中约有36.8%的样本未出现在采样集D'里。

7.1.2 基本特性

标准数据集应针对机器学习模型及系统的任务和场景特点，其基本特性包含数据量级、数据质量、数据分布等子特性。

a) 数据量级：应覆盖不同量级，数据量级应该能够适当反映模型目标应用的实际数据量级，应包含3个以上量级。根据不同任务，数据数量分为：
   1) 少：数据量较少，难以满足训练需求；
   2) 一般：数据量尚可，基本满足训练需求；
   3) 多：数据量充足，完全满足训练需求。
b) 数据质量：应充分考虑机器学习模型应用场景的真实情况，覆盖数据精度、数据噪声、数据缺失等情况，并应保证数据标注的质量。根据不同的任务分为：

1） 差：数据有大量噪音、缺失或相关度较低，难以满足训练需求；
2） 中：数据有部分噪音或缺失，基本满足训练要求；
3） 高：数据准确，完全满足训练要求。

c） 数据分布:样本分布应该考虑样本和标注的均衡分布,数据应从样本真实分布中独立采样生成，样本标注类别分布与任务真实分布应保持一致。根据不同的任务分为：
1） 差：数据有大量噪音、缺失或相关度较低，难以满足训练需求；
2） 中：数据有部分噪音或缺失，基本满足训练要求；
3） 高：数据准确，完全满足训练要求。

可用以下指标评估训练集、验证集和测试集的分布见表 2。

表2 标准数据集指标及测试方法

指标名称	指标描述	测试公式	测试值说明
群体稳定性(PSI)	标准数据集中每个变量在训练集,验证集和测试集中的分布稳定性。	$PSI=\sum_{i=1}^{B}(\hat{p}_i-\hat{q}_i)\times\ln\frac{\hat{p}_i}{\hat{q}_i}$ 式中： i——数据的第i个分组； B——总分组数； $\hat{p}_i$——验证或测试第i个分组占比； $\hat{q}_i$——训练第i个分组占比。	$\mathrm{PSI}<0.1$：指标稳定性很高； $0.1\le\mathrm{PSI}<0.25$：指标稳定性一般； $\mathrm{PSI}\ge0.25$：指标稳定性差。
均值稳定性	标准数据集中每个变量在训练集,验证集和测试集中的均值稳定性。	均值：$u=\frac{\sum_{i=1}^{n}x_i}{n}$ 式中： x_i——第i个样本； n——样本数目。	对比分析训练/验证/测试集指标的均值，差异越小越好。
方差稳定性	标准数据集中每个变量在训练集,验证集和测试集中的方差稳定性。	方差：$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-u)^2}{n-1}$ 式中： x_i——第i个样本； u——均值； n——样本数目。	对比分析训练/验证/测试集指标的方差，差异越小越好。

表 2　数据集指标及测试方法（续）

指标名称	指标描述	测试公式	测试值说明
偏度稳定性	标准数据集中每个变量在训练集，验证集和测试集中的偏度稳定性。	偏度：$S=\frac{\mathrm{E}(\mathrm{x_i}-u)^3}{\sigma^3}$ 式中： $\mathrm{x_i}$——第i个样本； u——均值； σ^3——三阶标准差。	$S=0$：正态分布； $S>0$：波形有右侧长尾； $S<0$：波形有左侧长尾； 差异越小越好。
峰度稳定性	标准数据集中每个变量在训练集，验证集和测试集中的峰度稳定性。	峰度：$\mathrm{K}=\frac{\mathrm{E}(\mathrm{x_i}-u)^4}{\sigma^4}$ 式中： $\mathrm{x_i}$——第i个样本； u——均值； σ^4——四阶标准差。	$\mathrm{K}=3$：正态分布； $\mathrm{K}>3$：波形平坦； $\mathrm{K}<3$：波形突兀； 差异越小越好。

7.2　功能性

7.2.1　正确性

采用测试集对机器学习模型及系统功能进行测试，测试机器学习模型及系统的数据精度的满足性、模型设计的正确性、代码实现的正确性、计算结果的正确性等。针对不同任务（如下，包含但不限于），主要依据数据集基准，采用不同测试方式进行测试：

a)　回归任务：采用机器学习模型预测结果的平均绝对误差（Mean Absolute Error）、均方误差（Mean Square Error）、均方根误差（Root Mean Square Error）等参数进行测试，具体参数选择视具体应用场景决定。回归任务的测试指标及方法见表 3；

表 3　回归任务的测试指标及方法

指标名称	指标描述	测试公式	测试值说明
平均绝对误差	所有单个观测值与算术平均值的偏差的绝对值的平均	$MAE=\frac{1}{n}\sum_{i=1}^{n}\lvert f_i-y_i\rvert$ 式中： MAE ——平均绝对误差； n ——测试样本个数； f_i ——第i个样本的模型预测值； y_i ——第i个样本的真实值。	$MAE\geq 0$，且 MAE 值越小越好。
均方误差	观测值与真值偏差平方和的平均值	$MSE=\frac{1}{n}\sum_{i=1}^{n}(f_i-y_i)^2$ 式中： MSE ——均方误差； n ——测试样本个数； f_i ——第i个样本的模型预测值； y_i ——第i个样本的真实值。	$MAE\geq 0$，且 MAE 值越小越好。
均方根误差	标准误差，是观测值与真值偏差的平方与观测次数比值的平方根。	$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i-y_i)^2}$ 式中： $RMSE$ ——均方根误差； n ——测试样本个数； f_i ——第i个样本的模型预测值； y_i ——第i个样本的真实值。	$RMSE\geq 0$，且 $RMSE$ 值越小越好。

b) 检索任务：采用机器学习模型检索结果的精确率（Precision）和召回率（Recall）。还可采用 F1 值、ROC 曲线、均值平均精度（Mean Average Precision）等进行综合评估。检索任务的测试指标及方法见表 4；

表 4　检索任务测试指标及方法

指标名称	指标描述	测试公式	测试值说明
精确率	对于给定的数据集，预测为真正例的样本占预测为正例的样本的比率。	$P=\frac{TP}{TP+FP}$ 式中： P ——精确率； TP ——预测为真正例的样本数； FP ——预测为假正例的样本数。	$0\le P\le 1$，且 P 值越大越好。
召回率	对于给定的数据集，预测为真正例的样本占所有实际为正例样本的比率。	$R=\frac{TP}{TP+FN}$ 式中： R ——召回率； TP ——预测为真正例的样本数； FN ——预测为假负例的样本数。	$0\le R\le 1$，且 R 值越大越好。
$F1$ 值	精确率和召回率的调和平均。	$F1=\frac{2*P*R}{(P+R)}$ 式中： $F1$ —— $F1$ 值； P ——精确率； R ——召回率。	$0\le F1$，且 $F1$ 值越大越好。

c)　分类任务：采用机器学习模型及系统分类结果的准确率（Accuracy）和错误率（Error　Rate）等进行测试，也采用 ROC 曲线等进行综合评估。其测试指标及方法见表 5；

表 5　分类任务的测试指标及方法

指标名称	指标描述	测试公式	测试值说明
准确率	对于给定的数据集，预测正确的样本占总样本的比率。	$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$ 式中： $Accuracy$ ——准确率； TP ——预测为真正例的样本数； TN ——预测为真负例的样本数； FP ——预测为假正例的样本数； FN ——预测为假负例的样本数。	$0\le Accuracy\le 1$，且 $Accuracy$ 越大越好。

表 5　分类任务的测试指标及方法（续）

指标名称	指标描述	测试公式	测试值说明
错误率	对于给定的数据集，预测错误的样本占总样本的比率。	$Error\ rate = 1 - Accuracy$ 式中： $Error\ rate$——错误率； $Accuracy$——准确率。	$0 \le Error\ rate \le 1$， $Error\ rate$越小越好。

d) 对于主观评价类型的任务（图像生成、语音合成等），需一定数量的观察者（专家、技术人员、普通用户等）在给定的观察条件下对机器学习模型产生的结果进行主观评价（评分、排序、比较等），并综合评价结果测试模型的正确性。

7.2.2　完备性

依据功能需求，采用测试集对机器学习模型及系统功能进行测试，检测机器学习模型及系统输出结果，测试功能实现与需求覆盖比、实现功能正交性。

7.2.3　恰当性

采用训练集对机器学习模型进行训练，通过使用不同规模的训练数据、改变训练数据的精度、设置不同的模型参数、采用不同的优化方法等方式，测试机器学习模型及系统的数据处理恰当性、模型设计恰当性、优化算法恰当性、模型实现恰当性、参数设置恰当性、训练操作恰当性等。

7.2.4　可迁移性

采用不同规模的训练数据集对机器学习模型进行训练和测试，评估其在不同规模下能否稳定迁移并实现预期功能（见6.2.1），测试其可扩展性；采用同领域不同数据集及不同领域的不同数据集对机器学习模型进行训练和测试，评估模型性能变化（见6.2.1），测试其可迁移性。

7.3　可靠性

7.3.1　鲁棒性

采用对抗攻击、噪声污染等手段生成对抗样本、噪声数据等对机器学习模型及系统进行测试，判断其能否正常运行并返回满足约定格式的运算结果。评价机器学习模型及系统能否正常运行、性能指标变化情况（见6.2.1）以及失效情况下恢复能力，测试危害检出性、抗攻击性、抗干扰性等。抗对抗性指标及测试方法见表6。

表 6　抗对抗性指标及测试方法

指标名称	指标描述	测试公式	测试值说明
抗对抗性	该指标为相对指标，用于多模型的鲁棒性对比评估。通过在验证集上增加噪声和离群点（建议可以对20%的验证样本加高斯噪声）对模型进行验证，然后计算模型拟合优度的变化值，变化值越小模型的抗对抗性越好。	$p=\lvert p_1-p_2\rvert$ 式中： p_1——未噪声和离群点前的模型拟合优度； p_2——为加入离群点后的模型拟合优度。 注：监督学习的模型拟合优度可以用准确率来测试。非监督学习的模型拟合优度可以用纯度测试。 $purity(\Omega,C)=\frac{1}{N}\sum_k \max_j \lvert w_k \mathrm{I}\ c_j\rvert$ $\Omega=\{w_1,w_2,\Lambda\ ,w_k\}$为样本聚类的集合，$C=\{c_1,c_2,\Lambda\ ,c_j\}$为样本分类集合； 式中： w_k——表示第k个聚类的集合； c_j——第j个分类； N——样本总数。	$p_1>0$； $p_2>0$； $p>0$； 且p值越小越好。

7.3.2　容错性

对机器学习模型及系统的数据输入、软硬件运行环境等进行极端或异常干扰。干扰方式包括运行资源受限性干扰，如强制限制内存使用量、降低中央处理器运行频率、限制处理器核心数量等；硬件失效性干扰，如数据采集设备强制失效等；对需要保存状态的机器学习模型，采用强制删除状态文件等方式进行干扰。测试机器学习模型及系统能否提示或内部消化在系统层出现的错误，及机器学习模型及系统正常运行的能力，测试模型系统的失效的避免性、误操作的抵御性、误操作的危害性等，其指标及测试方法见表7。

表 7　容错性指标及测试方法

指标名称	指标性质	测试公式	测试值说明
失效的避免性	避免关键和严重的失效比率	$X=\frac{A}{B}$ 式中： A——执行指标的测试要点的测试用例时，未发生关键的和严重的失效的测试用例数； B——指标的测试要点的测试用例总数。	$0.0\le X\le 1.0$ 越接近 1.0 越好，越能避免关键或严重的失效。

表 7　容错性指标及测试方法（续）

指标名称	指标性质	测试公式	测试值说明
误操作的抵御性	有效预防误操作的比率	$X=\frac{A}{B}$ 式中： A——执行误操作的测试要点的测试用例时，模型有效抵御该误操作的用例数； B——误操作的测试要点的测试用例总数。	$0.0\le X\le 1.0$ 越接近 1.0 越好，越能避免关键或严重的失效。
误操作的危害性	操作失效中未引起危害的比率	$X=1-\frac{A}{B}$ 式中: A——执行误操作的测试要点时发生无法抵御或未抵御时发生关键和严重危害的测试用例数。 B——执行误操作的测试要点的测试用例总数。	$0.0\le X\le 1.0$ 越接近 1.0 越好，越能避免关键或严重的失效。

7.3.3　易恢复性

采用软件故障、硬件故障检测机器学习模型及系统在软硬件出错环境下，修复软件系统和数据所需的平均时间、重新启动的平均时间、恢复的正确性（见6.2.1）、平均失效时间等，测试平均恢复时间、易重新启动性、易复原性、复原的有效性等见表8。

表 8　易恢复性指标及测试方法

指标名称	指标描述	测试公式	测试值说明
平均恢复时间	从模型失效到完全恢复所花费的平均时间。	$X=\frac{T_1+T_2+\Lambda+T_N}{N}$ 式中： $T(i=1,2,\Lambda,n)$——第 i 次模型恢复需要的时间； N——观察到的模型在故障中进入恢复的总次数。	$0<X$ 越小越好。
易重新启动性	在一定的试验周期内，从模型失效到模型正常启动所花费的平均时间。	$X=\frac{T}{N}$ 式中： T——总的重启时间； N——观察到的模型失效时间。	$0<X$ 越小越好，模型失效时间越短。

表 8　易恢复性指标及测试方法（续）

指标名称	指标描述	测试公式	测试值说明
易复原性	在异常情况下需要修复模型的成功比率。	$X=\frac{A}{B}$ 式中： A——模型成功完成恢复的测试用例数； B——执行的模型恢复测试用例总数。	$0<X<1.0$ 越接近 1.0 越好，模型更容易修复
复原的有效性	模型失效后的修复能力有效程度。	$X=\frac{A}{B}$ 式中： A——满足模型恢复时间的成功完成恢复的测试用例数； B——执行的模型恢复测试用例总数。	$0<X<1.0$ 越接近 1.0 越好，模型更容易修复。

7.4　效率

7.4.1　时间消耗

时间消耗指机器学习模型训练的计算量或计算复杂度等。在指定实验条件下，对机器学习模型及系统，采用不同规模的训练集进行多次训练，统计不同数据量下训练收敛的平均耗时、训练单轮的平均耗时以及运行时间随数据规模变化的曲线。

时间消耗也指机器学习模型测试的计算量或计算复杂度等。在指定实验条件下，对机器学习模型及系统进行多次测试，统计单个数据处理的平均耗时。

7.4.2　资源消耗

资源消耗指机器学习模型训练的存储量或空间算复杂度（以字节为单位）等。在指定实验条件下，对机器学习模型及系统，采用不同规模的训练集，进行多次系统训练，统计不同数据量下训练过程的最大存储（内存/显存/硬盘等）占用、平均存储占用、最大带宽（硬盘吞吐、网络流量等）占用、平均带宽占用、最大计算资源（CPU、GPU等）占用、平均计算资源占用等以及上述指标随数据规模变化的曲线。

资源消耗也指机器学习模型测试的存储量或空间算复杂度（以Byte为单位）等。在指定实验条件下，对机器学习模型及系统，进行多次系统测试，统计单个数据处理的平均最大存储占用、平均存储占用、最大带宽占用、平均带宽占用、最大计算资源占用、平均计算资源占用等。资源消耗也采用模型测试的存储量或空间复杂度等方式评价。

7.5　维护性

7.5.1　规范性

机器学习模型及系统的设计和开发应遵循机器学习模型相关的标准或规范，依据机器学习模型及系统约定的规范，测试机器学习模型及系统设计的规范性、模型训练的规范性、模型测试的规范性等。

机器学习模型及系统的设计和开发应遵循软件产品维护相关的标准或规范，依据软件开发规范以及模型开发的平台、语言规范，采用软件测试方法，测试机器学习模型及系统设计、开发、运行和维护中命名规范性、模型封装性、版本兼容性等，测试机器学习模型及系统代码的易读性、系统版本兼容性等。

7.5.2 收敛性

采用标准训练数据集测试机器学习模型的收敛性，采用不同规模、不同类型训练集对模型进行训练，对模型训练的稳定性、收敛时间和收敛值等测试元进行测试（见6.4.1）。具体指标包括不同场景下训练收敛迭代次数、训练时间、模型损失等。

7.5.3 易改变性

依据软件开发规范以及开发的平台、语言规范，检查变更说明文档的完整性、模块间的耦合性；采用软件测试方法，测试阶段机器学习模型及系统的缺陷定位以及训练阶段中机器学习模型及系统修改、模型系统扩充的成本，测试机器学习模型及系统的变更说明文档的完整性、模块间的耦合性、变更模块的可验证性等。

7.5.4 可验证性

采用形式化验证方法等测试模型计算过程的可验证性；针对机器学习模型的应用场景和目标，采用关联分析、因果分析等手段分析模型计算结果的意义，测试模型计算结果的可解释性。

采用软件形式化验证、软件测试等手段测试机器学习模型及系统功能的可验证性等。

附 录 A
（资料性附录）

A.1 机器学习模型及系统测试实例

A.1.1 实例：癫痫会话诊断测试

表 A.1 癫痫会话诊断的深度学习模型及系统质量测试表

	质量指标 （文件里的测试元）	基本指标描述 （详细的测试元）	测试方法	达标的基本要求
数据集	数据集	数据规模	会话数据量	训练集>500 小时，测试集>50 小时
		数据质量	音频采样率	>=8kHz
功能性	完备性	功能实现与需求覆盖比	查看神经网络能否输出被试分类诊断信息	能够对被试会话特征定量评分并作出诊断分类
		实现功能正交性	无	无
	正确性	算法设计的正确性	分析算法基本流程，检测神经网络输出分类是否符合算法设计要求	符合算法设计要求
		计算的正确性	Accuracy 指标	>90%
		数据精度的满足性	检测网络能否正常训练，数值计算是否溢出	训练正常
		代码实现的正确性	检测网络训练是否收敛、网络输出是否正常	训练收敛、输出正常
		参数调整正确性	采用不同参数，检查神经网络输出情况	达到精度要求
	可迁移性	同领域数据集可迁移性	采用同一病种内不同数据集，检查神经网络输出情况	训练收敛、输出正常
		领域可迁移性	采用不同病种亚型数据集，检查神经网络输出情况	训练收敛、输出正常
	可解释性	模型原理的可解释性	模型的原理和搭建过程可逻辑阐述并提交伦理委员会专家审查	专家审查通过
		模型运算流程的可解释性	模型计算过程的各流程分支可逻辑阐述并提交伦理委员会专家审查	专家审查通过
		模型计算结果的可解释性	模型计算结果可通过关联分析、因果分析、统计模型等手段阐述模型计算结果的临床意义	专家审查通过
可靠性	鲁棒性	抗干扰性	采用噪声污染和数据隐藏方法、混叠方法生成噪声和非规范性数据，检查神经网络输出情况	训练收敛、输出正常

表 A.1　癫痫会话诊断的深度学习模型及系统质量测试表（续）

	质量指标 （文件里的测试元）	基本指标描述 （详细的测试元）	测试方法	达标的基本要求
		抗攻击性	采用对抗攻击方法生成对抗样本，检查神经网络输出情况	训练收敛、输出正常
效率	时间消耗	模型收敛时间	检查模型收敛所需时间	<5 小时
		模型训练单轮时间	检查模型所需单轮训练时间	<20 分钟
	资源消耗	模型执行所需显存大小	确认所需最低计算机显存配置	>=5GB
维护性	易改变性	文档的完整性	检查代码与开发文档描述的一一对应性	文档覆盖度>90%
	收敛性	收敛时间	采用不同规模、不同癫痫病种亚型训练集对模型进行训练，对模型训练的收敛时间进行测试。	<30 分钟

A.1.2　实例：目标跟踪测试

表 A.2　目标跟踪的深度学习模型及系统质量测试表

	质量指标 （文件里的测试元）	基本指标描述 （详细的测试元）	测试方法	达标的基本要求
数据集	数据集	数据规模	数据集视频数据量	训练集>1000，测试集>100
		数据质量	视频质量	单个视频时长>15s，分辨率>480x320
		数据注释	每帧图像中的目标数量，视频中的目标编号数量，目标的标注框	每帧图像中的目标数量>1，目标编号数量>1000，目标的标注框刚好把完整的目标标出
功能性	正确性	数据精度的满足性	检测网络能否正常训练，数值计算是否溢出	训练正常
		模型设计的正确性	分析算法基本流程，检测神经网络输出分类是否符合算法设计要求	符合算法设计要求
		代码实现的正确性	检测网络训练是否收敛、网络输出是否正常	训练收敛、输出正常
		计算结果的正确性	跟踪准确性 MOTA	>50%
			跟踪稳定性 MOTP	>60%
	完备性	功能实现与需求覆盖比	查看模型能否输出目标在每一帧图像下的位置	能够输出目标在每一帧图像下的位置
		实现功能正交性	无	无

表 A.2　目标跟踪的深度学习模型及系统质量测试表（续）

	质量指标 （文件里的测试元）	基本指标描述 （详细的测试元）	测试方法	达标的基本要求
	恰当性	模型的数据处理恰当性	无	无
		模型设计恰当性	无	无
		优化算法恰当性	无	无
		模型实现恰当性	无	无
		参数设置恰当性	无	无
		训练操作恰当性	无	无
可靠性	可迁移性	同领域数据集可迁移性	采用不同场景下同类目标的跟踪数据集，检查神经网络输出情况	训练收敛，输出正常
		领域可迁移性	采用不同类目标的跟踪数据集，检查神经网络输出情况	训练收敛，输出正常
	鲁棒性	抗干扰性	在数据中添加噪声，检查神经网络的输出情况	训练收敛，输出正常
	容错性	无	无	无
	易恢复性	无	无	无
效率	时间消耗	模型测试时间	跟踪平均每帧耗时	<10ms
	资源消耗	资源占用	确认所需最低计算机显存配置	>=5GB
维护性	规范性	无	无	无
	收敛性	无	无	无
	易改变性	无	无	无
	可验证性	无	无	无

中国电子工业标准化技术协会（CESA）是全国电子信息产业标准化组织和标准化工作者自愿组成的社会团体。广泛联系全国电子信息产业标准化机构和标准化工作者，协助政府部门搞好电子信息产业标准化工作，开拓信息技术领域的标准化工作是中国电子工业标准化技术协会的主要工作内容之一。中国境内从事科研开发、制造、营销和服务的企事业单位、高等院校、社会组织和个人均可随时向中国电子工业标准化技术协会团体标准工作部提出团体标准项目建议。

中国电子工业标准化技术协会标准按照《电子工业标准化技术协会协会团体标准管理办法》进行制定和管理。

在本标准实施过程中，如发现需要修改或补充之处，请将意见和有关资料寄至中国电子工业标准化技术协会，以便修订时参考。

中国电子工业标准化技术协会地址：北京市海淀区万寿路27号
电话：010 - 64102952 电子邮箱： standards@cesa.cn
网址：www.cesa.cn