ICS 国际标准分类号
CCS 中国标准文献分类号

团　　体　　标　　准

T/CES XXX-XXXX

电力人工智能感存算一体化系统测试方法

Test method of power AI systems with integrated capabilities of sensing, data storage, and processing

（征求意见稿）

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会　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》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。

本文件由国网信息通信产业集团有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：国网信息通信产业集团有限公司、福建亿榕信息技术有限公司、中国科学院上海微系统与信息技术研究所。

本文件主要起草人：李强、赵峰、王秋琳、庄莉、李建华、梁懿、王营冠、何为、李炳森、邱镇、卜智勇、宋立华、郑耀松、丘志强、冯小蔚、琚诚、陈江海、吕志超、王燕蓉、张维、王婧。

本文件为首次发布。

电力人工智能感存算一体化系统测试方法

1 范围

本文件规定了面向输电、变电、配电等电力领域的人工智能感存算一体化系统测试方法，可对基于人工智能的系统功能、典型应用场景性能进行评估，评测场景包括图像分类、目标检测、目标分割、目标识别、目标跟踪、语音识别、文字识别、本地语音唤醒、负荷辨识等。

本文件适用于生产厂商、研究机构、应用厂商及第三方机构对具备感存算一体化系统进行测试评估，也适用于生产厂商对感存算一体化系统的设计。

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 1.1-2020 标准化工作导则 第1部分：基本术语

GB/T 9813.2-2016 计算机通用规范 第2部分：便携式微型计算机

GB/T 9813.3-2017 计算机通用规范 第3部分：服务器

GB/T 5271.28-2001 信息技术 词汇 第28部分：人工智能 基本概念与专家系统

GB/T 5271.34-2006 信息技术 词汇 第34部分：人工智能 神经网络

GB/T25000.51-2016 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

GB/T 36572-2018 电力监控系统网络安全防护导则

GB/T 26866-2022 电力时间同步系统检测规范

T/CES 128-2022 电力人工智能平台总体架构及技术要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

神经网络模型 neural network model

神经网络的抽象模型，它能用软件来模拟或作为神经计算机加以实现。

[来源：GB/T 5271.34-2006,34.01.10]

3.2

训练 training

教会神经网络在输入值的样本和正确输出值之间做出结合的步骤。

[来源：GB/T 5271.34-2006,34.03.18]

3.3

推理 inference

从已知前提导出结论的推理方法。

注1：在人工智能领域，前提是事实或者规则。

注2：术语“推理”既指过程也指结果。

[来源：GB/T 5271.28-2001,28.03.01]

3.4

训练集 training set

数据集的子集，用于训练模型。

3.5

测试集 test set
数据集的子集，用于在模型经由验证集的初步验证之后测试模型。

3.6
批次 batch
模型训练的一次迭代（即一次梯度更新）中使用的样本集。

3.7
批次大小 batch size
一个批次中的样本数。批次大小在训练和推理期间通常是固定的。

3.8
轮次 epoch
使用训练集的全部数据对神经网络模型进行一次完整的训练，被称之为一代训练。

3.9
作业到达 workload
一组被一同送入训练或推理系统的 N 个样本，N 为正整数。

4 符号、代号和缩略语

下列符号、代号和缩略语适用于本文件。
SUT：被测系统（System Under Test）
mAP：平均精度均值（Mean Average Precision）
mIoU：平均交并比（Mean Intersection Over Union）
FPS：每秒帧率（Frame Per Second）
FAR：误识率（False Accept Rate）
FRR：拒识率（False Reject Rate）
IR：识别正确率（Identification Rate）
WER：词错误率（Word Error Rate）
SER：句错误率（Sentence Error Rate）
NNM 神经网络模型（Neural Network Model）
Training：训练
Inference：推理

5 测试说明

5.1 测试对象

本文件的测试对象具体形式有以下两种：

a） 含有计算机视觉感存算一体化系统的控制主机，指以卡/棒等形态进行使用的感存算一体智能系统，如 GPU、FPGA、ASIC 等感存算一体模块，可通过 PCIE、USB 等接口与测试主机连接；

b） 搭载人工智能处理器的感存算一体模块。

5.2 测试内容

感存算一体化系统的测评指标，主要包括基本技术规格、功能、性能、电力应用场景测试等部分，在依据本文件进行测试的过程中：

a） 涉及功能、性能、电力应用场景等相关指标将通过第三方测试工具进行评测；

b） 涉及基本技术规格的指标将采信被测对象标称值及其他技术信息，作为先进性的参考。

5.3 环境要求

除另有规定外，环境应满足 GB/T 9813.2-2016 或 GB/T 9813.3-2017 中大气条件的规定，其中：

a） 温度：5℃~35℃或 15℃~35℃；

b） 相对数度湿度：25%~75%；

c） 大气压：86kPa~106kPa。

此外，若送测方有更为严苛的要求，应满足送测方提出的温度、湿度等。

5.3 基本要求

基本要求包括且不限于下述内容：

a） 应支持至少一种存算一体技术，包括但不限于查存计算、近存计算、存内计算、存内逻辑等；

b） 支持主流的人工智能框架：TensorFlow、Pytorch、Caffe/Caffe2、Mxnet、ONNX、MindSpore（昇思）或 PaddlePaddle（飞桨）等；

c） 模型精度：FP64、FP32、FP16、INT4、INT8、INT16、BP16 或混合精度等。其中，训练场景精度应支持 FP16、FP32、FP64，推理场景下精度应支持 INT8、FP16；

d） 应支持以下至少 1 种自主可控加速器，包括但不限于昇腾、智芯等；ASIC 类的加速器，如 NPU 等；FPGA 类型的加速器；GPU 类型的加速器；

e） 控制主机处理器架构：X86 架构、ARM、RISC-V 或 MIPS 等架构；

f） 电力人工智能模型应满足电力业务应用场景所需的计算、算力资源等；

g） 测试用仪器设备均应经过计量部门检定合格，并在有效期内，专用测试设备必须经过严格标定，并在标定有效期内使用。

5.4 测试过程

测试过程包括：

a） 测试申请：由送测单位提供测试委托书，申请对样品进行测试；

b） 制定测试大纲：依据本文件与实际测试需求制定测试方案，确定测试内容，各项测试的进度安排，资源要求，测试资料，测试工具，系统的配置方式，回归测试的规定等以及评价标准，如果无法构建出要求相同的测试环境，后续需进一步分析由于测试环境与使用环境不一致所带来对测试结果的影响,形成测试大纲；

c） 样品送测：由送测单位送测样品；

d） 测试环境部署：根据送测样品部署相应测试环境；

e） 测试类型：

功能测试。在构建的测试环境下，对样品覆盖的功能进行测试，检验各测试项目是否实现、是否正确实现。

性能测试。在构建的测试环境下，进行实时监测和数据收集，利用准备好的测试数据集对被测系统进行各测试项目测试，分为正常情况、人为设置的系统资源紧缺异常情况、人为设置的高负载高负荷情况，即将测试数据集一次输入被测系统，并按照被测系统的使用方法开展测试活动，检验各测试项是否达标、是否能够保持；

回归测试。定期、不定期测试进行回归测试，对被测系统进行重新功能和性能测试，确认每次更新和迭代修改后的系统仍满足规定的要求。

f） 出具报告：完成测试后，收集整合测试数据，对测试结果进行汇总、深入分析和综合评价，形成测试报告。

5.5 场景信息

电力人工智能测试场景包括：

a） 典型电力人工智能应用场景：计算机视觉任务，包括图像分类、目标检测、目标分割、目标识别、目标跟踪、文字识别等，语音识别、本地语音唤醒等任务，电力专用应用负荷辨识等任务，与其对应电力生产环节见电力人工智能感存算一体化系统设计规范附表 A.1；

b） 数据集：公开数据集或真实电力应用场景数据集；

c） 模型：经典神经网络模型或自定义神经网络模型。

5.6 作业到达

电力应用场景作业到达的方式包含以下几种模式：

a） 单路模式：测试主机向被测系统串行发送作业请求，单次作业请求包含 1 个样本，被测系统完成单次作业运算返回结果得到测试主机确认之后，测试主机再向被测系统发送下一条作业请求，并以此循环；

b） 测试主机发送新的作业：如果被测系统已经及时完成上一次的作业运算并返回结果，则被测主机按照限定延迟间隔发送一个新的作业请求。如果被测系统未能及时完成，则新的请求被丢弃并被记为一次作业超时。

c） 云服务模式：作业到达被测系统服从泊松分布：

$$P(k,\lambda)=\frac{\lambda^k e^{-\lambda}}{k!}$$

其中，k 表示在某单位时间内到达的作业数，λ 表示单位时间内平均作业平均到达次数。每次作业可含有多个样本，每次含有的样本数量 Y，Y 服从正态分布：

$$Y \square N(\mu,\sigma^2)$$

其中，μ 为样本数量均值，σ 实际到达样本数量的离散程度。

d） 本地模式：所有作业一次性全部到达被测系统。

6 功能测试

6.1 接入管理测试

验证感存算一体化系统的接入管理能力：

a）检查感存算一体化系统支持 RS485、RS232 等串口通信接口以及模拟量、开关量、数字量等信号接口各类冯·诺依曼架构设备、感存算一体化设备通过串口、以太网接入。

b）检查感存算一体化系统支持至少一种视频、 图像、语音、文本等类型感知数据接入。

c）检查感存算一体化系统支持通过 GB/T 28181、ONVIF、RTSP 等接入不同厂商的视频监控摄像机。

d）检查感存算一体化系统支持 H264/H265 等主流编解码协议，支持存储视频调阅回看等功能。

e）检查感存算一体化系统支持 NSA、SA 的 5G 全频段接入（含大网及专网）。

6.2 网络测试

a）感存算一体化系统解析蓝牙、Zigbee、Wifi、LoRa、NFC、RFID 等无线或有线一种或多种传输协议，满足各网络协议接入要求。

b）具备数据转发功能的感存算一体化系统，测试主机通过被测系统进行数据转发，数据应能被正确转发到指定通信接口。

c）感存算一体化系通过协议传输数据，检查支持 MQTT、HTTP/HTTPS、DL/T 698.45、CoAP、DL/T645、IEC60870-5-104、IEC61850 等多种通信协议。

d）感存算一体化系接入 SDN、TSN 等新型网络设备/系统，检查是否支持 SDN、TSN 等新型网络。

6.3 计算及存储测试

计算及存储测试见表 1。

表 1 计算存储测试

序号	测试项目	测试内容	测试准则
1	异构计算架构	人工智能感存算一体化系统对异构计算架构的支持	查看被测系统软件清单中的异构计算架构支持情况，与真实情况符合。 如果支持则通过，否则为不通过。

2	存算一体技术	人工智能感存算一体化系统对存算一体技术的支持	查看被测系统软件清单中的存算一体技术支持情况，与真实情况符合。 如果支持则通过，否则为不通过。
3	本地化存储	本地化存储能力	被测系统连接测试仪表，验证具备本地化存储能力。
4	多类型数据存储和处理能力	多类型数据存储和结构化、非结构化数据处理能力	在被测系统接入其他设备感知数据，添加数据规则引擎，判断处理后的结果和预设相符。 如果相符则通过，否则为不通过。
5	指令集和计算单元协同技术	不同类型指令集和不同体系架构计算单元协同技术的支持	查看被测系统软件清单以及代码的不同类型指令集和不同体系架构计算单元协同技术支持情况，与真实情况符合。 如果支持则通过，否则为不通过。
6	多类型人工智能平台	开放集成多种 AI 训练和推理平台，兼容多厂商计算单元	查看被测系统软件清单以及代码的多种 AI 训练和推理平台，兼容多厂商计算单元的支持情况，与真实情况符合。 如果支持则通过，否则为不通过。

6.4 AI 能力测试

AI 能力测试见下表，对模型推理性能测试描述详见 7 AI 模型推理测试。

表 2 AI 能力测试

序号	测试项目	测试内容	测试准则
1	处理器	人工智能感存算一体化系统对处理器类型的支持	查看被测系统硬件清单中的设备基本信息，包括支持的处理器类型、加速器类型以及精度支持情况，与真实情况符合。 如果支持则通过，否则为不通过。
2	加速器	人工智能感存算一体化系统对加速器类型的支持	
3	精度类型	人工智能感存算一体化系统对精度类型的支持	
4	AI 模型部署	人工智能感存算一体化系统对部署 AI 模型的支持	登录测试系统，部署 AI 模型，通过在线测试方式验证 AI 模型可提供正常 AI 服务，验证 AI 模型功能。
5	AI 服务管理	人工智能感存算一体化系统对部署 AI 服务管理的支持	登录测试系统，选择一个正在运行的 AI 服务，执行查看和停止操作，验证 AI 服务可管理功能。
6	AI 模型库管理	人工智能感存算一体化系统对 AI 模型库管理的支持	登录测试系统，执行 AI 模型导入和删除操作，验证 AI 模型库管理功能。
7	人工智能框架	人工智能感存算一体化系统对人工智能框架的支持	查看受测设备软件清单中的人工智能支持情况，与现实信息符合。 如果支持则通过，否则为不通过。

8	模型训练	使用的人工智能框架中模型训练支持功能	查看被测系统支持的人工智能框架技术规格书或开源网站。 如果支持则通过，否则为不通过。
9	模型推理	使用的人工智能框架中模型推理支持功能	

6.5 模型推理测试

模型推理测试详情见“7.AI 模型推理测试”。

6.6 远程维护测试

a）验证时间同步符合 GB/T 26866-2022 的要求。

b）登录测试系统，执行被测系统软件安装，验证是否支持软件安装功能。

c）登录测试系统，执行被测系统远程固件升级，验证是否支持远程固件升级。

d）检查日志，如果日志中记录了日志类型、登录时间、登录地址、登录用户名、开启或停止服务等远程维护操作则评测通过，否则评测不通过。

6.7 自治能力测试

a）登录测试系统，模拟感存算一体化系统外部网络故障，验证故障不会影响电力应用提供服务。

b）登录测试系统，恢复感存算一体化系统外部网络故障，验证故障恢复不会影响电力应用提供服务。

6.8 可扩展性测试

a）验证感存算一体化系统支持提供对 DAS、NAS（CIFS，NFS）的访问；

b）验证感存算一体化系统支持对主流 CentOS 、Ubuntu 等 Linux、windows 平台的访问；

c）扩展存储空间，验证感存算一体化系统支持存储空间动态扩展。

6.9 安全性测试

a）验证信息安全符合 GB/T 36572-2018 的要求。

b）验证数据部分符合《中华人民共和国数据安全法》的规定。

7 AI 模型推理测试

7.1 测试指标

主要测试指标见表 3：

表 3 典型电力应用场景测试指标

类 型	测试指标	
感存算一体化系统	训练指标（非必需）	训练时间
		训练能耗
	推理指标	最大吞吐性能
		平均前向推理速率
		前向推理时延
		功耗
		能效比
		模型推理准确度

7.2 测试指标描述

7.2.1 训练时间

在 5.3 技术要求规定下，训练某一神经网络达到指定精度所需要的时间。

7.2.2 训练能耗

在 5.3 技术要求规定下，训练某一神经网络达到指定精度时被测系统的能耗。

7.2.3 最大吞吐性能

指被测系统在训练过程或前向推理过程（包括预处理，后处理）中可同时处理的最大样本数量。

7.2.4 平均前向推理速度

指被测系统在指定 batch 下，在单位时间内使用神经网络模型完成测试数据集运算的平均样本数量。

7.2.5 前向推理时延

被测系统在指定 batch 下前向推理运算过程（不包括预处理，后处理）中，根据作业到达中的方式，计算从样本输入被测系统完成到计算结果由被测系统开始输出间的时间间隔。

7.2.6 功耗

在 5.3 技术要求规定下，未运行推理运算时被测系统的静态功耗；以及进行指定模型推理运算过程中被测系统的平均功耗。可以带有控制主机的功耗。

7.2.7 能效比

模型推理过程中，被测系统在单位时间内执行作业的次数与感存算一体化系统功耗之比。

7.2.8 模型推理准确度

指定任务场景下，被测系统使用某一神经网络模型完成测试后，所得到的平均模型前向推理准确度数值，不同电力应用场景下采用的模型推理的评价指标不同，详见 7 测试场景。

7.3 训练测试方法

训练过程中，记录必要的测评数据：

a） 训练时间：不包含训练过程中使用测试集，测量当前模型准确率及准确率门限的比较时间；

b） 训练次数：对同一目标模型的训练过程，重复训练的次数计数；

c） 单次训练时间：记录单次训练过程的总体训练时间（不含每 epoch 后模型准确率计算时间）；

d） 单次训练能耗：记录单次训练过程的总体训练能耗（不含每 epoch 后模型准确率计算能耗）。

训练结果数据如下：

a） 平均训练时间： 平均训练时间=$\sum$ 每次训练的时间/次数；

b） 平均训练能耗：平均训练能耗=$\sum$ 每次训练的能耗/次数；

c） 实际准确率：模型训练过程中，在测试集上的准确率。

7.4 推理测试方法

7.4.1 最大吞吐性能

在前向推理计算前，通过不断调整 batch 参数，增加单位时间内输入样本的数量，最终确定被测系统的最大吞吐性能。

7.4.2 平均前向推理速率

单位：FPS（处理图像数量/秒）等。

平均前向推理速率=总测试样本数量/总预测时间

注：

总测试样本数量指测试数据集中样本的总数量。

总预测时间指被测系统使用神经网络模型时，在测试数据集上完成测试所使用的总计算时间。

7.4.3 前向推理时延

单位：ms（毫秒）。

前向推理时延=当前样本处理结果开始输出的时刻-当前样本输入完成时刻

7.4.4 功耗

单位：w(瓦)。

a） 针对板卡或棒形态的被测系统，对被测系统在执行测试期间的功耗情况进行记录，最终通过计算得到该硬件的平均功耗情况。

功耗值=基准电流差 X 电压

b） 针对服务器形态的被测系统，测试被测系统在任务处理阶段总的能耗与所用时间的比值，得到平均功率，再测一个不进行深度学习任务处理的平均功率，计算两个平均功率之差即为被测系统的功耗。

7.4.5 能效比

单位：FPS/W（处理图片数量/秒/瓦特）等。

能效比=平均前向推理速率/功耗

7.4.6 模型推理精度

针对不同电力应用场景，包括 top-1、top-5 算法准确率、mAP、mIoU、F-Score、FAR、FRR 等。

8 典型应用场景测试

8.1 图像分类场景

a） 模型：ResNet-50、ResNet-101、VGG16、Inception-v3 和 MobileNet v2；

b） 数据集：ImageNet；

c） 性能评价指标：

Top-1 准确率：图像分类结果中排名第一的分类类别与实际结果相同的准确率；

Top-5 准确率：图像分类结果中排名前五的分类类别与实际结果相同的准确率。

8.2 目标检测场景

a） 模型：两阶段模型（Faster R-CNN、Mask R-CNN）和一阶段模型（YOLO、SSD）；

b） 数据集：Microsoft COCO、CPLID、OPDL；

c） 性能评价指标：

平均精度均值（Mean average precision，mAP）：数据集中所有类别的平均精度的均值。

平均精度均值=所有类别平均精度值之和/所有类别的数目。

8.3 目标分割场景

a） 模型：Deeplab v3++、DeepMask、Fast-SCNN；

b） 数据集：TTPLA、PLD-UAV、VOC 2012、CityScapes；

c） 性能评价指标：

平均交并比（Mean intersection over union，mIoU）:所有类别的交并比 IoU 的平均值。

每个类别的交并比 IoU 为真实值（ground truth）和预测值（predicted segmentation）两个集合的交集合并集之比。

交并比（Intersection over union，IoU）：检测结果的矩形框与样本标注的矩形框的交集与并集的比值。

F 分数：综合衡量精确率和召回率的指标。

$$\text{F-Score}=\frac{\text{Precision}\times\text{Recall}}{\beta^2\times\text{Precision}+\text{Recall}}$$

其中，

精确率（Precision）：识别正确的结果在所识别的结果中所占的比率；

召回率（Recall）：识别正确的结果占数据集中所有要识别出的总数的比率。

8.4 目标识别场景

a） 模型：FaceNet、Object Recognition、DeepID3、ArcFace；

b） 数据集：WebFace、LFW；

c） 性能评价指标：

误识率（False accept rate，FAR）：将其他目标误作指定目标的概率；

拒识率（False reject rate，FRR）：将指定目标误作其他目标的概率；

识别正确率（Identification rate）：正确识别目标次数与参与识别目标的总次数之比。

8.5 目标跟踪场景

a） 模型：UDT、TADT、UMA Tracker；

b） 数据集：MOT16、VOT、OTB；

c） 性能评价指标：

准确率（Accuracy）：跟踪器在单个测试序列下的平均重叠率（两矩形框的相交部分面积除以两矩形框的相并部分的面积）；

鲁棒性（Robustness）：单个测试序列下的跟踪器失败次数，当重叠率为 0 时即可判定为失败；

平均重叠期望（Expect average overlap rate，EAO）：对每个跟踪器在一个短时图像序列上的非重置重叠的期望值；

成功率曲线下面积，单目标跟踪任务时，模型准确率为纵轴，1 减去准确率的值为横轴，绘制成功率曲线，计算曲线下面积，为将正样本判断为正样本的可能性大于判断为负样本的可能性的概率；

多目标跟踪的准确度（Multiple object tracking accuracy，MOTA）：体现在确定目标的个数，以及有关目标的相关属性方面的准确度，用于统计在跟踪中的误差积累情况；

多目标跟踪的精确度（Multiple object tracking precision，MOTP）：体现在确定目标位置上的精确度，用于衡量目标位置确定的精确程度。

8.6 行为检测场景

a） 模型：TGM、PGCN、SSN、BSN、BMN；

b） 数据集：ActivityNet、Kinetics、THUMOS14、AVA、CASIA、UCSD Ped2、ShanghaiTech；

c） 性能评价指标：

视频帧准确度平均值：各类标记结果正确的视频帧数占标记结果中该类视频帧总数的比例均值，即视频帧中正确标签的第 i 类视频帧数量与检测结果中第 i 类视频帧数比值的均值。

视频片段准确率均值：各类标记结果正确的视频片段数占标记结果中该类视频总数的比例均值，即视频片段正确标签的第 i 类视频片段数量与检测结果中第 i 类视频片段数比值的均值。

8.7 语音识别场景

a） 模型：DeepSpeech2；

b） 数据集：AISHELL-2；

c） 性能评价指标：

词错误率（WER）：语音识别后，识别词错误唤醒的次数占总唤醒的百分比；
句错误率（SER）：语音识别后，识别句子错误唤醒的次数占总唤醒的百分比。

8.8 文字识别场景

a） 模型：CRNN、CPTN；
b） 数据集：MJ_LMDB、IIIT5k 等；
c） 性能评价指标：
F 分数：综合衡量精确率和召回率的指标。

$$\text{F-Score}=\frac{\text{Precision}\times\text{Recall}}{\beta^2\times\text{Precision}+\text{Recall}}$$

其中，
精确率（Precision）：识别正确的结果在所识别的结果中所占的比率；
召回率（Recall）：识别正确的结果占数据集中所有要识别出的总数的比率。

8.9 本地语音唤醒场景

a） 模型：DFSMN、CNN、DNN；
b） 数据集：Speech commands dataset；
c） 性能评价指标：
词错误率（WER）：语音识别后，识别词错误唤醒的次数占总唤醒的百分比；
句错误率（SER）：语音识别后，识别句子错误唤醒的次数占总唤醒的百分比。

8.10 负荷辨识场景

a） 模型：DAE、ShortSeq2Point、WindowGRU；
b） 数据集：REDD、BLUED 等；
c） 性能评价指标：
精确率（Precision）：TP/（TP+FP），识别正确的结果在所识别的结果中所占的比率；
正确率（Accuracy）：（TP+TN）/（TP+FP+TN+FN）
True positives（TP）：被正确划分为正例的个数，即实际为正例且被分类器划分为正例的样本数；
False positives（FP）：被错误划分为正例的个数，即实际为负例且被分类器划分为正例的样本数；
False negatives（FN）：被错误划分为负例的个数，即实际为正例且被分类器划分为负例的样本数；
True negatives（TN）：被正确划分为负例的个数，即实际为负例且被分类器划分为负例的样本数。

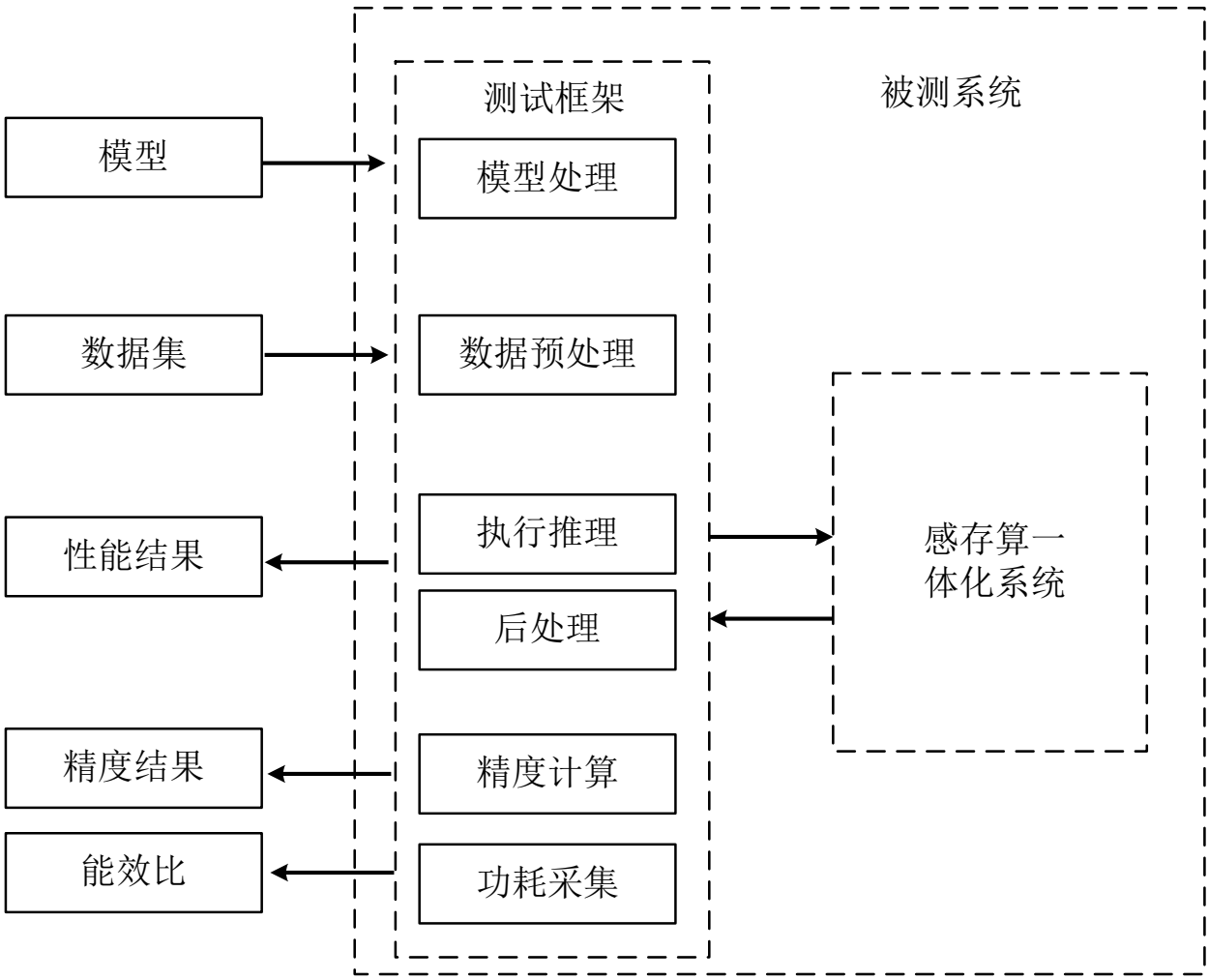
/

附 录 A
（规范性附录）

A.1 测试框架

A.1.1 推理测试框架

推理测试框架见图 A.1。

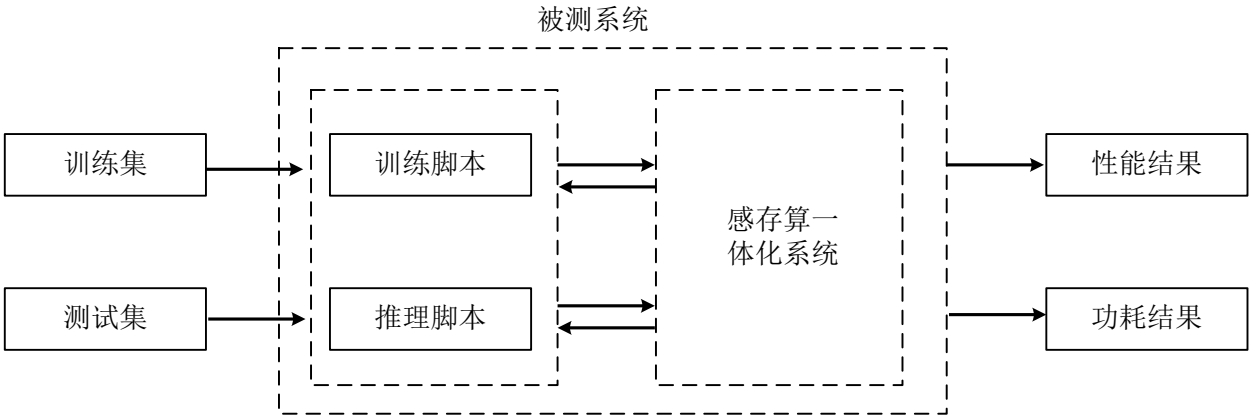


1 输入:本标准指定的模型和数据集;
2 模型处理，可选，包括量化、使用模型转换工具对模型进行转换;
3 送测方提供使能其系统的接口供测试框架调用，包括初始化、模型加载、模型执行和模型卸载。

图 A.1 推理测试框架

A.1.2 训练测试框架

训练测试框架说明见图 A.2



1 输入：训练集用于模型训练；测试集用于精度验证。
2 推理脚本，可以单独写，也可以作为分支写在训练脚本中，可以每 N 个 step 后推理-一次，也可以最后推理一次
3 训练性能统计的前提是精度达标。即:达到精度指标时模型训练所用的时间

图 A.2　训练测试框架（可选）

参 考 文 献

[1] GB/T25000.51-2016 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

[2] 周志华 机器学习 清华大学出版社 2016

[3] GB/T 5271.34-2006 信息技术 词汇 第 34 部分：人工智能 神经网络

肖进胜, 申梦瑶, 江明俊, 雷俊峰, 包振宇.融合包注意力机制的监控视频异常行为检测.自动化学报,2022,48(12):2951-2959