

SF

中华人民共和国司法行政行业标准

SF/T 0182—2024

硅藻检验技术规范 基于人工智能技术的
自动化方法

Technical specification for diatom testing—Automated method based on artificial
intelligence technology

2024 - 12 - 30 发布

2025 - 06 - 01 实施

中华人民共和国司法部 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	1
5 方法原理	1
6 仪器设备	1
7 检验步骤	2
参考文献	7

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。

本文件由司法鉴定科学研究院提出。

本文件由司法部信息中心归口。

本文件起草单位：司法鉴定科学研究院、上海市公安局、南京医科大学、中国医科大学、山西医科大学、广州市公安局。

本文件主要起草人：黄平、张吉、王亚辉、马开军、陈峰、赵锐、孙俊红、赵建、秦志强、张建华、杨明真、陈敏。

硅藻检验技术规范 基于人工智能技术的自动化方法

1 范围

本文件规定了基于人工智能技术的法医学硅藻检验自动化方法，包括方法原理、仪器设备和检验步骤。

本文件适用于由人工智能模型完成的法医学尸体器官组织、体液及水样硅藻的定性、定量及分类检验。

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。

GA/T 813 人体组织器官中硅藻硝酸破机法检验

GA/T 1662 法庭科学 硅藻检验技术规范 微波消解-真空抽滤-显微镜法

3 术语和定义

下列术语和定义适用于本文件。

3.1

硅藻 diatom

具有硅质细胞壁的水生单细胞藻类。

注：粒径一般为数微米至数百微米。

4 缩略语

下列缩略语适用于本文件。

AUC：曲线下面积值（Area Under the Curve）

CPU：中央处理器（Central Processing Unit）

CUDA：计算统一设备架构（Compute Unified Device Architecture）

DPI：每英寸点数（Dots Per Inch）

HOG：方向梯度直方图（Histogram of Oriented Gradient）

JPG：联合图片组（Joint Picture Group）

LBP：局部二值模式（Local Binary Pattern）

PNG：便携式网络图像（Portable Network Graphics）

TIF：标签图像文件格式（Tag Image File Format）

5 方法原理

通过数字化病理学技术将实体样本底物转换成高分辨率图像数据；应用人工智能算法对数字化图像中的硅藻区域进行自动化识别和科属分类；根据人工智能算法统计的硅藻数量及科属分类结果并结合人工确认，完成法医学硅藻的定性、定量及分类检验工作。

注：实体样本底物包括组织、体液及水样涂片以及滤膜等。

6 仪器设备

6.1 显微扫描或拍照系统

- 6.1.1 显微扫描或拍照系统可在放大倍率不小于 400 倍的视场下获取整张或局部样本底物的高清图像，且具备实时自动对焦功能。
- 6.1.2 所获得的高清图像文件可通过显微图像分析软件或程序进行处理，如切割、均一化和标准化等。
- 6.1.3 处理后的高清图像可以 JPG、TIF 或 PNG 等常规格式保存，图像分辨率应不低于 72 DPI。

6.2 数字计算机

- 6.2.1 数字计算机应不低于以下配置。
 - a) 运行内存 8GB。
 - b) CPU 处理器 1 个：主频 2.3GHz、八核心、十六线程、L3 缓存 11MB、动态加速频率 3GHz。
 - c) 显卡 1 个：显存容量 11GB、显存位宽为 352bit、显存频率为 14000MHz，且支持 CUDA 运算。
- 6.2.2 数字计算机宜具备以下配置。
 - a) 运行内存 128GB。
 - b) CPU 处理器 2 个：主频 2.5GHz、二十四核心、四十八线程、L3 缓存 20MB、动态加速频率 3.4GHz。
 - c) 显卡 2 个：显存容量 11GB、显存位宽为 352bit、显存频率为 14000MHz，支持 CUDA 运算。

7 检验步骤

7.1 人工智能模型构建

7.1.1 数据库构建

7.1.1.1 基本要求

数据库应包含识别数据集和分类数据集。其中，识别数据集是由含有完整或部分硅藻的图像样本和无硅藻的底物背景图像样本组成的数据集；分类数据集是由含有不同硅藻科属图像样本组成的数据集。数据库类型分为外部数据库和内部数据库。

7.1.1.2 外部数据库

建立外部数据库的前提是本地实验室所采用的预处理方法，可以有效降低不同数据源之间图像样本的差异性。外部数据库主要用于模型的训练和训练过程中的效能评价，本地实验室仪器设备获取的图像样本应作为模型在硅藻识别和分类方面的外部评价数据，具体如下：

- a) 外部数据库来源具有多样性，可来自实验室仪器平台、外部实验室、网络数据库及商业数据库等；
- b) 外部数据库中的图像样本应进行像素大小调整，使图像大小符合人工智能模型输入规格；
- c) 外部数据库中的图像样本应进行像素预处理（如均一化和灰度化等），降低不同数据源所带来的像素差异性，提高模型的泛化能力。

7.1.1.3 内部数据库

若不采用图像预处理或所使用的图像预处理方法难以有效降低不同数据源之间图像样本的差异性，本地实验室则应根据自身仪器型号和样本制备方法等条件构建数据库，用于模型的训练和评价，具体如下：

- a) 可使用不同型号仪器、样本制备方法或网络数据库等获取的图像样本用于人工智能模型的预训练，但不宜使用这些图像样本来评价本地实验室仪器平台模型在硅藻识别和分类方面的真实场景效能；
- b) 应针对本地实验室仪器平台单独构建数据库，用于评价模型在硅藻识别和分类方面的真实场景效能；
- c) 若本地实验室更换仪器设备、更改显微扫描或拍照系统的参数以及硅藻样本制备方法，则应基于在用设备和方法重新建立数据库用于模型的训练和评价。原数据库可用于最新模型的预训练；

- d) 获取的图像样本可进行像素大小调整,使图像大小符合人工智能模型的输入规格,可选择对图像样本像素进行预处理(如均一化和灰度化等),用于降低同一数据源中各图像样本之间的像素差异性,提高模型的泛化能力。

7.1.1.4 数据库图像样本

- 7.1.1.4.1 在数据库单张图像样本中,目标硅藻区域与实际完整硅藻面积比例应不低于30%。
7.1.1.4.2 数据库图像样本可根据背景及不同科属硅藻分成多组数据集,每组数据集包含原始图像应不少于5000张,且每张图像样本拥有相应分组标签信息。
7.1.1.4.3 各组数据集图像数量应将组间数量差距保持在20%以内。

7.1.2 人工智能模型训练

7.1.2.1 训练目的

人工智能模型的训练目的在于针对特定(如识别和分类)任务利用数据库中的大量数据样本对智能化模型进行模式特征学习。经学习后的模型可用于训练数据以外的同类型、同分布数据的准确、客观预测。人工智能模型的特定任务取决于数据库的性质。其中,识别数据集主要用于训练将硅藻与背景杂质相区分的模型架构,而分类数据集则用于训练区分不同科属硅藻的模型架构。

7.1.2.2 训练方法

- 7.1.2.2.1 在基于端到端的模型训练中,所采用的模型架构(如深度学习模型)直接从图像样本中提取相关特征信息用于模型内部参数的拟合和校正。
7.1.2.2.2 基于分治法的模型训练应借助额外方法先提取图像中的特征信息,形成向量数据后才可用于模型(如一般机器学习模型)内部参数的拟合和校正。

7.1.2.2.3 基于端到端的模型训练

基于端到端的模型训练应在符合6.2规定的数字计算机上完成,具体如下。

- a) 采用的人工智能模型算法应涉及卷积神经网络、目标识别算法以及其他可用于图像特征自动化提取的模型架构。
- b) 应将数据库图像样本按适当比例设置训练、验证和测试样本集。其中:
 - 1) 训练样本集、验证样本集和测试样本集中的数据分布应保持一致;
 - 2) 训练样本集和验证样本集分别用于模型的训练和训练过程中的模型效能监测和评估,测试样本集则用于模型训练完毕后的性能泛化指标评价;
 - 3) 图像样本来自于不同个体或实体底物样本,为防止数据泄漏,来自于同一个体或实体底物样本中的图像样本应被分配到同一样本集中。
- c) 可对训练样本集中的图像样本进行扩增,包括图像翻转、图像旋转、随机调整图像对比度及亮度等。
- d) 将扩增后的7.1.2.2.3 c) 图像样本用于模型的迭代训练,每次迭代获得的模型对验证样本集中的图像样本进行预测。
- e) 根据模型验证样本预测结果及其对应的真实分组标签,计算目标损失函数值用于评估每次迭代的模型预测效能。
- f) 当目标损失函数值随着迭代次数增加而趋于稳定时,为防止模型过拟合的出现,可在目标损失函数值趋于稳定的迭代临界点后10次~50次获取相应最优模型。

7.1.2.2.4 基于分治法的模型训练

基于分治法的模型训练应在符合6.2规定的数字计算机上完成,具体如下。

- a) 采用的人工智能模型算法应涉及一般机器学习模型(如偏最小二乘法、支持向量机、随机森林和感知器神经网络等)以及其他不能直接处理图像矩阵数据的模型架构。
- b) 应将数据库图像样本按适当比例设置训练、验证和测试样本集。其中:
 - 1) 训练样本集、验证样本集和测试样本集中的数据分布应保持一致;

- 2) 训练样本集和验证样本集用于模型的训练和训练过程中的模型效能监测和评估，测试样本集则用于模型训练完毕后的性能泛化指标评价；
- 3) 图像样本来自于不同个体或实体底物样本，为防止数据泄漏，来自于同一个体或实体底物样本中的图像样本应被分配到相同数据集中。
- c) 可对训练样本集中的图像样本进行扩增，包括图像翻转、图像旋转、随机调整图像对比度及亮度等。
- d) 通过特征提取方法（如深度学习特征提取模型、HOG 特征提取算法和 LBP 特征提取算法等）获取图像样本中的特征数据并将这些特征数据整合成向量数据。每个图像样本的向量数据维度及特征排列次序应保持相同。
- e) 如特征提取方法涉及深度学习模型，其内部参数可通过训练和验证图像样本并结合特定损失函数予以事先拟合和校正。
- f) 训练样本的向量数据应用于人工智能模型架构的参数拟合，同时结合网格搜索方法确认模型最佳超参数组合并获得候选模型。
- g) 应用候选模型预测验证样集中图像样本的向量数据并计算分类准确率，准确率高于 0.9 时可将其确认为最优模型。

7.1.3 人工智能模型效能评价

7.1.3.1 效能评价指标

7.1.3.1.1 查准率

对于给定的数据集，正确预测的正例样本与被预测为正例的样本比率，见公式（1）。

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \dots\dots\dots (1)$$

式中：

- TP ——实际是正例且被预测为正例的样本数量；
- FP ——实际是负例但被预测为正例的样本数量；
- precision——查准率。

7.1.3.1.2 查全率

对于给定的数据集，正确预测的正例样本与实际为正例样本的比率，见公式（2）。

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \dots\dots\dots (2)$$

式中：

- TP ——实际是正例且被预测为正例的样本数量；
- FN ——实际是正例但被预测为负例的样本数量；
- recall——查全率。

7.1.3.1.3 准确率

对于给定的数据集，正确分类的样本数占总样本数的比率，见公式（3）。

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{ALL}} \dots\dots\dots (3)$$

式中：

- TP ——实际是正例且被预测为正例的样本数量；
- TN ——实际是负例且被预测为负例的样本数量；
- ALL ——所有样本的数量；
- accuracy——准确率。

7.1.3.1.4 F1 分数

模型查准率和查全率的一种加权平均，也是一种综合评价指标，见公式（4）。

$$F1\text{score} = 2 \times \frac{\text{recall} \times \text{precision}}{(\text{recall} + \text{precision})} \dots\dots\dots (4)$$

式中：

recall ——查全率；

precision——查准率；

F1score ——F1分数。

7.1.3.1.5 AUC

表示预测正例排在负例前面的概率，用于评估模型的分类效能。

7.1.3.2 模型的内部评价

模型的内部评价应在符合6.2规定的数字计算机上完成，具体如下：

- a) 应用最优模型对测试样本集进行预测。如采用基于分治法的模型架构，测试样本集应进行特征提取形成向量数据后才可输入至人工智能模型。特征提取方法应与模型训练时采用的特征提取方法相一致；
- b) 应根据测试样本集预测结果及其对应的真实分组标签，计算准确率、查全率、查准率、F1 分数及 AUC 值；
- c) 最优模型在测试样本集预测中，其准确率、查全率、查准率、F1 分数及 AUC 值均应大于 0.9。

7.1.3.3 模型的外部评价

模型的外部评价应在符合6.2规定的数字计算机上完成，具体如下：

- a) 在本地实验室仪器设备针对不同器官、体液及水样分别制作 10 份底物样本(如组织样本涂片)，每份底物样本中的实际硅藻数量不低于 100 个，科属类别不低于 5 种。底物样本中的实际硅藻数量和科属分类由本实验室 2 名及以上，日常从事硅藻检验工作，且具有一定分类经验的技术人员予以确认；
- b) 按照 7.2.2 的规定在本地实验平台将底物样本进行数字化转化；
- c) 按照 7.2.3 和 7.2.4 的规定应用最优模型对数字化底物样本进行硅藻定性、定量及科属分类；
- d) 人工智能模型识别正确的硅藻数量应占底物样本实际硅藻数量的 80%以上；
- e) 人工智能模型的硅藻科属分类平均准确率应在 90%以上。

7.2 自动化硅藻分析的具体实施

7.2.1 硅藻检验样本制备

按照GA/T 813、GA/T 1662以及其他可用于肺、肝、肾、骨髓、体液及水样等样本的硅藻检验方法，所获得的硅藻检验样本可用于7.2.2的步骤。

7.2.2 硅藻检验样本数字化转换

硅藻检验样本数字化转换应在符合6.1规定的显微扫描或拍照系统上完成，具体步骤如下。

- a) 将获取的检测底物样本（如组织样本涂片）放入显微扫描或拍照系统的载物台上。
- b) 确定扫描区域，可设置自动聚焦选项。
- c) 在 400 倍及其以上倍率的视场下获取检测底物样本上扫描区域的数据图像，数据图像符合：
 - 1) 数据图像应大于或等于人工智能模型输入大小规格；
 - 2) 数据图像表示分辨率应不低于 72 DPI；
 - 3) 色彩深度为本地实验室仪器设备实际参数。

7.2.3 应用人工智能识别模型对数据图像中硅藻区域定位

在符合6.2规定的数字计算机上应用人工智能识别模型对数据图像中硅藻区域定位，具体步骤如下：

- a) 将 7.2.2 步骤中获取的数据图像进行图像处理，采用自动切割方法形成多张子图像。所生成的子图像大小应与人工智能模型输入的规格相符；

- b) 可选择对 7.2.3 a) 中的子图像进行预处理（如均一化和灰度化等），预处理方法应与 7.1.1.2 c) 或 7.1.1.3 d) 图像样本所使用的预处理方法相同；
- c) 如采用基于分治法的模型架构，子图像样本应进行特征提取形成向量数据后才可输入至人工智能识别模型。特征提取方法应与 7.1.2.4 d) 图像样本所使用的特征提取方法相同；
- d) 经人工智能识别模型预测，可确认所检测的子图像中是否含有硅藻（完整或部分硅藻）并记录该张子图像在数据图像上的坐标信息。

7.2.4 应用人工智能分类模型对数据图像中硅藻分类

在符合6.2规定的数字计算机上应用人工智能分类模型对数据图像中硅藻分类，具体步骤如下：

- a) 若子图像含有因切割操作造成的不完整硅藻，可根据 7.2.3 d) 中获取的坐标信息在数据图像上截取相同大小的子图像，该子图像应包含有完整硅藻；
- b) 样本在消化、离心和震荡等过程中造成硅藻破坏时，如对应子图像中包含的硅藻碎片可通过形状或纹理予以确认，则按 7.2.4 a) 进行处理；
- c) 可选择对 7.2.4 a) 中的子图像进行预处理（如均一化、灰度化等），预处理方法应与 7.1.1.2 c) 或 7.1.1.3 d) 图像样本所使用的预处理方法相同；
- d) 如采用基于分治法的模型架构，截取的子图像样本应进行特征提取形成向量数据后才可输入至人工智能分类模型。特征提取方法应与模型训练时采用的方法相一致；
- e) 经人工智能分类模型预测，确认截取子图像中硅藻的形态学科属并记录该子图像在数据图像上的坐标信息。

7.2.5 人工确认

在符合6.2规定的数字计算机上，检验人员对人工智能识别模型和分类模型筛选的含有硅藻的子图像进行人工校对并获取对应子图像在数据图像中的坐标信息。

7.2.6 结果呈现

在符合6.2规定的数字计算机上结果呈现如下：

- a) 根据人工智能模型和人工确认结果，对样本中的硅藻进行定性分析和定量分析，并统计不同科属硅藻的数量；
- b) 根据含有硅藻的子图像坐标信息，在数据图像中截取相应大小代表性的图像用于检验报告附图。

参 考 文 献

- [1] GB/T 5271.1—2000 信息技术 词汇 第1部分：基本术语
 - [2] AIOSS—01—2018 人工智能 深度学习算法评估规范
 - [3] J Zhang, DN Vieira, Q Cheng, et al. DiatomNet v1.0: a novel approach for automatic diatom testing for drowning diagnosis in forensically biomedical application[J]. *Comput Meth Prog Bio*, 2023, 232: 107434
 - [4] J Zhang, YY Zhou, DNo Vieira, et al. An efficient method for building a database of diatom populations for drowning site inference using a deep learning algorithm[J]. *Int J Legal Med*, 2021, 135: 817-827
 - [5] YY Zhou, J Zhang, J Huang, et al. Digital whole-slide image analysis for automated diatom test in forensic cases of drowning using a convolutional neural network algorithm[J]. *Forensic Sci Int*, 2019, 302: 109922.
 - [6] B Gloria, D Oscar, P Anibal, et al. Automated Diatom Classification (Part A): Handcrafted Feature Approaches[J]. *Appl Sci*, 2017, 7:753
 - [7] P. Anibal, B. Gloria, D. Oscar, et al. Automated diatom classification (Part B): a deep learning approach[J]. *Appl Sci*, 2017, 7:460
 - [8] M Kloster, D Langenkemper, M Zurowietz, et al. Deep learning-based diatom taxonomy on virtual slides[J]. *Sci. Rep*, 2020, 10:14416
 - [9] J Salido, C Sánchez, J Ruiz-Santaquiteria, et al. Automated digital microscopy platform for automatic identification of diatoms[J]. *Appl Sci*, 2020, 10:6033
 - [10] WM Yu, Y Xue, R Knoops, et al. Automated diatom searching in the digital scanning electron microscopy images of drowning cases using the deep neural networks[J]. *Int. J. Legal Med*, 2021, 135:497-508
-