检测数据偏移下病理视觉语言模型的性能下降

Hao Guan，Li Zhou

arXiv:2601.00716v1 [cs.CV] 2 Jan 2026

摘要—视觉语言模型在医学图像分析和疾病诊断方面展现出巨大潜力。然而，模型部署后，如果输入数据分布偏离开发阶段所观察到的分布，其性能可能会下降。检测此类性能下降对于确保临床可靠性至关重要，但对于在无标签条件下运行的大型预训练视觉语言模型而言，仍然颇具挑战。本研究探讨了先进病理视觉语言模型在数据偏移下的性能下降检测。我们考察了输入层面数据偏移和输出层面预测行为，以了解二者在监测模型可靠性方面各自的作用。为便于系统分析输入数据偏移，我们开发了*DomainSAT*，这是一款轻量级工具箱，配有图形界面，集成了具有代表性的偏移检测算法，并支持直观地探索数据偏移。我们的分析表明，尽管输入数据偏移检测能够有效识别分布变化并提供早期诊断信号，但它并不总是与实际性能下降相对应。基于这一观察，我们进一步研究了基于输出的监测，并提出一种无标签的基于置信度的退化指标，可直接捕捉模型预测置信度的变化。我们发现，该指标与性能下降密切相关，可有效补充输入偏移检测。在一个用于肿瘤分类的大规模病理学数据集上开展的实验表明，结合输入数据偏移检测与基于输出置信度的指标，能够更可靠地检测和解释视觉语言模型在数据偏移下的性能下降。这些发现为监测数字病理学中基础模型的可靠性提供了一套实用且互补的框架。

索引词—数据偏移检测、医学人工智能监测、性能下降、视觉-语言模型、数字病理学、科学软件、人工智能可靠性

I. 引言

RECENT 视觉语言模型（VLM）的进步极大拓展了人工智能在医学领域的能力 [1]–[4]。通过联合编码视觉和文本信息、视觉语言模型（例如 CLIP [1]）及其医学适配模型，使临床成像模态中的灵活零样本分类、图像-报告检索和图像描述生成成为可能。在数字病理学领域，视觉语言模型为病理学计算诊断提供了一个promising的基础。

H. Guan 和 L.Zhou 任职于布里格姆妇女医院普通内科与初级保健部、哈佛医学院医学系（美国马萨诸塞州波士顿，邮编 02115）。通讯作者：H. Guan（hguan6@bwh.harvard.edu）。

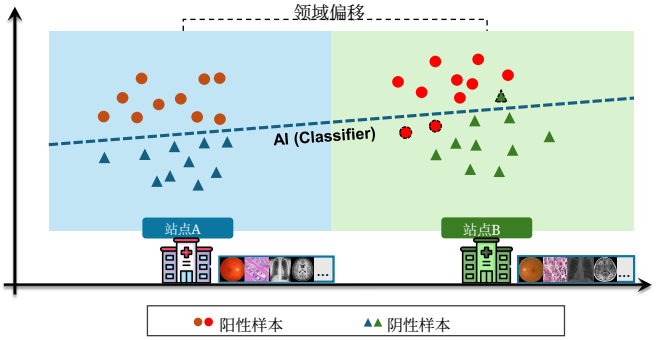


图 1。 不同医疗机构之间的领域（数据）偏移及其对医疗人工智能模型性能影响的概览。在站点A部署的人工智能模型运行良好，但应用于站点B时，该模型在数据偏移下出现性能下降。

尽管具有这些优势，视觉语言模型在临床部署中的长期可靠性仍不确定。部署后，这些模型可能会遇到与观测分布存在显著差异的输入数据，这些差异可能源于扫描仪、染色流程或数据采集地点的变化。这就是众所周知的领域（数据）偏移问题 [5]–[8]。如图 1所示，此类数据偏移可能导致性能下降，损害人工智能辅助诊断系统的安全性和可信度 [9]。因此，检测性能下降对于防范临床风险和指导模型维护至关重要。

然而，检测大型预训练病理视觉语言模型的性能下降面临着若干根本性挑战。首先，尽管视觉语言模型在医学成像中的应用迅速普及，但其部署后的可靠性，尤其是在真实世界数据偏移下的性能下降，尚未得到系统研究。其次，在实际医疗环境中，部署后通常无法获得真实标签，因此使用准确率或曲线下面积等标准指标直接检测性能下降并不可行。

为应对这一挑战，我们通过联合分析输入数据偏移和输出层面的预测，研究最先进的病理视觉语言模型在数据偏移下的性能下降。为系统地探索输入数据偏移，我们开发了**DomainSAT**，这是一款轻量级的基于图形用户界面的工具箱，集成了具有代表性的偏移检测算法，可直观地检查不同数据集中的数据偏移。我们的分析证实，输入级偏移检测在识别分布变化并提供预警信号方面发挥着重要作用，但也表明，仅凭输入数据偏移并不总能预测下游性能下降。

受这一观察启发，我们进一步考察输出层面的信号，并提出一种简单而有效的**无需标签的**基于置信度的性能退化指标。该指标能够捕捉数据条件发生偏移时模型预测置信度的变化。其基本直觉是，随着视觉语言模型的运行状态逐渐偏离其可靠工作范围，其输出置信度分布会变得更难区分且更不确定，这表明性能正在下降。

我们使用一个涵盖多个医疗机构且存在扫描仪引起的变异的大规模病理学数据集，对所提出的框架进行了评估。实验结果表明，基于输出的置信度指标能够紧密跟踪性能退化，而输入级偏移检测则提供了互补的诊断信息。综合这些信号，可以更可靠地监测和解释数据偏移条件下病理学视觉语言模型的性能退化。

本研究的主要贡献有以下三点：

- 对数据偏移导致的最先进病理视觉语言模型性能退化进行了深入研究。
- 提出了一种无标签的基于置信度的退化指标（CDI），无需真实标签即可为视觉语言模型退化提供有效信号。
- DomainSAT，这是一个轻量级图形用户界面工具箱，集成了具有代表性的偏移检测方法，并促进了对数据偏移模式的直观可视化。

这些组件为数字病理学基础模型的可靠性监测提供了实用框架，DomainSAT 工具箱可通过以下网址公开获取：https://github.com/guanharry/DomainSAT。

II. 相关工作

A. 视觉-语言模型

视觉语言模型（VLM）学习图像与文本的联合表征，从而实现零样本分类和检索。CLIP [1], BLIP-2 [10],和 LLaVA [11] 等通用视觉语言模型已展现出在多种视觉领域中的强大泛化能力。受其成功启发，研究人员开发了多种医学视觉语言模型，以改善临床结局。这些模型包括 MedCLIP [2], BiomedCLIP [12], 和 LLaVA-Med [3], ，它们利用配对的医学图像-文本数据来增强医学理解。

在数字病理学领域，近期出现了专用领域视觉语言模型，以应对高分辨率组织病理图像和细粒度组织语义所带来的独特挑战。值得注意的是，本研究使用的最先进模型 PathGen-CLIP [13]，是在 160 万个病理图像 - 文本对上训练而成。PathGen-CLIP 能够生成强大的病理学专用嵌入，并已在肿瘤分类任务中展现出优异的性能。尽管进展迅速，但病理视觉语言模型在真实世界数据偏移下的可靠性和性能下降仍鲜有研究，这正是本研究的动机。

B. 领域偏移检测

域（数据）偏移检测方法通常可分为三类：基于距离的、基于统计量的和基于机器学习的，如表I所示。

表 IDOMAINSAT 工具箱中包含的代表性领域偏移检测算法。

类别	算法	参考
基于距离的	MMD	[15]
	Wasserstein距离	[16]
	马氏距离	[17]
	JS 散度	[18]
	KL 散度	[19]
基于统计量的	Kolmogorov-Smirnov检验	[20]
	Wilcoxon秩和检验	[21]
	Cram´er-von Mises检验	[22]
	χ^2 检验	[23]
基于机器学习	域分类器	[24]
	C2ST（逻辑分类器）	[25]
	C2ST（随机森林）	[26]
	自动编码器	[27]

1) 基于距离的方法：这些方法使用距离度量量化参考数据与目标数据之间的差异，其中数值越大通常表示领域偏移越强，性能下降的风险也越高。在 [14],中，预训练自动编码器从胸部X光片中提取潜在特征，并计算最大均值差异距离，以检测细微的分布变化，包括与新出现的疾病（如新冠肺炎）相关的变化。Stacke *etal.* [6] 使用基于CNN提取的特征，并应用 Wasserstein 距离和 KL 散度来评估医学图像分析中的潜在空间偏移。

2) 基于机器学习的方法：偏移检测的一种常见策略是训练域分类器（例如逻辑回归），以区分源样本和目标样本 [24]。较高的分类准确率表明两者的分布存在显著差异。这种方法称为分类器双样本检验（C2ST） [28]–[30], ，广泛用于高维数据。例如， [31] 将C2ST应用于糖尿病视网膜病变筛查监测，以检测与图像质量、共病和人口统计特征相关的偏移。

3) 基于统计量的方法：这些方法应用假设检验来比较源分布和目标分布，通常返回p值，以便进行可解释的决策。它们通常用于一维特征，以确定偏移的来源。Kolmogorov-Smirnov（KS）检验 [20] 常用于数据偏移检测。CheXstray [23] 在DICOM元数据、图像特征和模型输出上采用KS检验，以进行实时监测。

III. 域偏移分析工具箱

A. 工具箱设计与工作流程

1) 总体：如图 2 所示，DomainSAT（**领域偏**移**分**析工具箱）由三个主要模块组成：a）数据加载模块，负责以标准化格式导入源数据集和目标数据集；b）算法模块，集成了经典的领域偏移检测方法，使用户能够进行深入分析；c）输出模块，用于生成易于理解的偏移检测结果以及信息丰富的可视化内容。

为最大限度提升易用性，该工具箱提供**图形用户界面**（**GUI**），支持离线探索性分析和回顾性分-

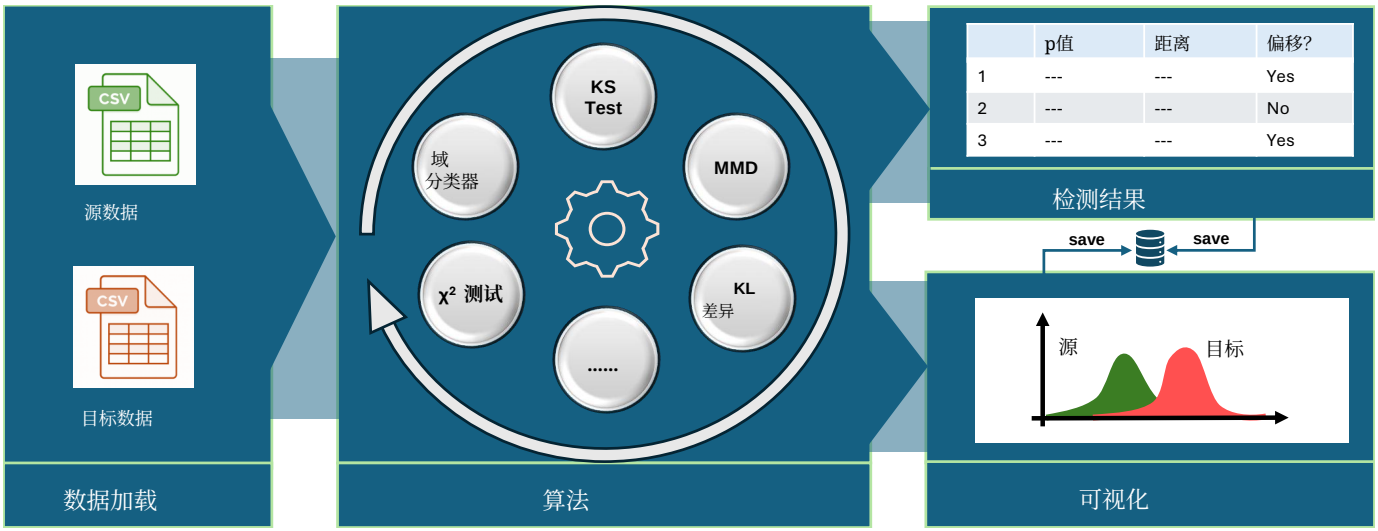


图2. DomainSAT 工作流概览。该工具箱包含三个主要组件：(1) 用于导入和预处理数据集的数据加载模块；(2) 集成多种领域偏移检测方法的算法模块；以及 (3) 用于生成检测结果和可视化内容的输出模块。

析、数据分布可视化以及领域偏移评估，无需编写任何代码。

2) 工作流程：用户通过图形用户界面上传源数据集和目标数据集（*.csv* 格式），即可开始操作。随后，用户可以从工具箱中选择一种或多种领域偏移检测算法。执行后，DomainSAT 会生成详细的输出，包括*p* 值、距离度量或分类器得分，这些结果可保存为*.csv* 文件以供进一步分析。此外，工具箱还提供可视化功能，用于展示源领域和目标领域中各项特征的分布。这些直观的可视化对比有助于深入分析领域偏移的特征级表现。

B. 算法

DomainSAT 工具箱集成了多种常用的领域偏移检测算法，详见表I。根据底层机制，这些算法分为三大类：基于距离的方法、统计检验方法和基于机器学习的方法。

基于距离的方法会计算一个数值距离分数，用以量化源数据集和目标数据集之间的差异。当该分数超过预设阈值时，通常会判定存在领域偏移。

统计检验方法用于评估源数据和目标数据是否来自相同的分布。较小的 *p* 值（通常低于 0.05 或 0.02）表示存在具有统计显著性的数据偏移。

基于机器学习的方法会训练一个分类器，以区分源数据和目标数据。如果分类器的性能显著优于随机猜测（通过准确率或 ROC曲线下面积（曲线下面积）等指标衡量），则表明存在可检测到的领域偏移。

该工具箱及其包含的所有算法均使用 Python 实现。作为完全开源的项目，用户可以遵循现有流水线，自由扩展工具箱，添加并测试新算法。所有已实现的方法均可跨平台运行（例如，Windows、macOS），且无需 GPU。

表 II本研究所用 CAMELYON17 站点汇总。站点 1–3 合并为分布内（ID）参考集，站点 4 和 5 则作为分布外（OOD）数据集。

数据集	分配	扫描仪	图像数量	阳性数
站点-1、2、3	ID（参考）	Pannoramic	33,560	50%
站点-4	分布外-S1（未见）	滨松	34,904	50%
站点-5	分布外-S2（未见）	飞利浦	85,054	50%

IV. 探究病理视觉语言模型的性能下降

在本节中，我们通过联合分析输入级偏移信号和输出级指标，研究病理学视觉语言模型在数据偏移下的性能下降，并强调二者在可靠的人工智能模型监控中所发挥的互补作用。

A. 数据集

1) 概述：我们使用 WILDS Camelyon17 数据集，这是一个公开可用的组织病理学基准，旨在研究领域偏移 [32], [33]。该数据集记录了五家医院（站点 1、2、3、4、5）之间因染色流程和扫描仪类型差异而产生的真实世界数据偏移。站点 1–3 使用相同的组织病理图像采集扫描仪，而站点 4 和 5 各自使用不同类型的扫描仪。该数据集由从全切片病理图像中提取的图像块组成，并为每个图像块提供二元标签，标示是否存在肿瘤。表 II总结了我们实验中使用的站点。

2) 参考站点和分布外站点：在本研究中，我们定义如下：

- 分布内（ID）/参考站点：将站点 1–3 合并，以模拟部署参考基准。这反映了模型开发期间可用的分布，并为监测提供基线。
- 分布外（OOD）站点：站点 4（OOD-S1）和站点 5（OOD-S2）作为未见数据，使用

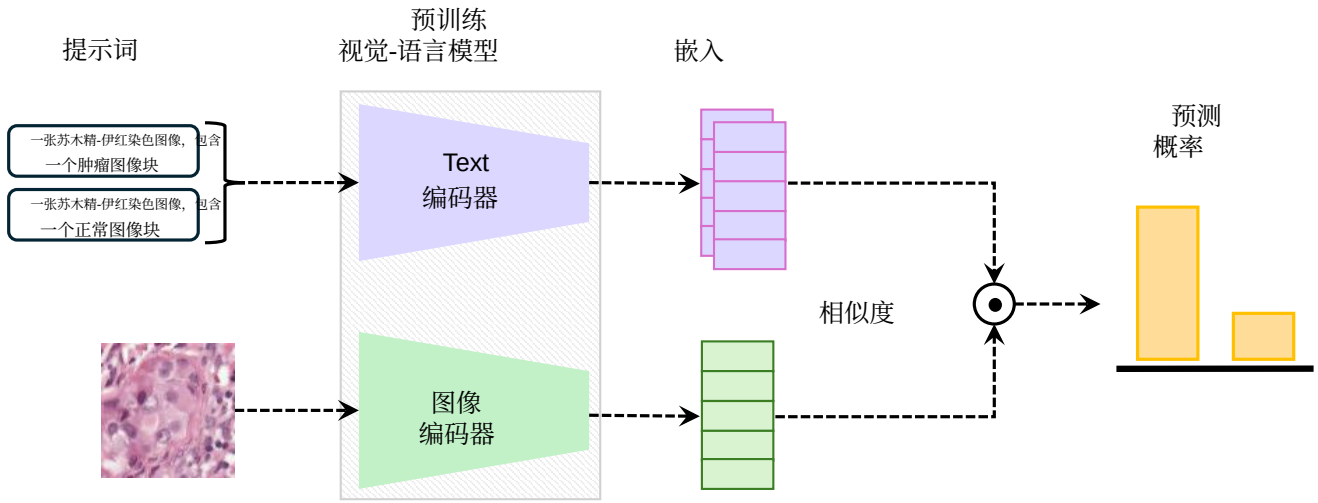


图 3. 预训练视觉语言模型（PathGen-CLIP）用于组织病理图像分类的示意图。输入包括一张组织病理图像和两条文本提示（“一张肿瘤图像块的苏木精-伊红染色图像”和“一张正常图像块的苏木精-伊红染色图像”）。图像和文本分别经过编码，并通过相似度评分进行比较。随后将得到的相似度分数归一化为两个类别（肿瘤和正常）的预测概率。

不同的扫描仪和染色条件采集。这些数据模拟了部署后遇到的真实领域偏移。

B. 模型

我们评估 PathGen-CLIP [13],这一用于肿瘤分类的先进病理学视觉-语言模型，它以 CLIP 为基础 [34]，采用 ViT-B/16 图像编码器和标准的基于Transformer的文本编码器。为模拟人工智能产品部署后的真实场景，所有实验中模型始终保持完全冻结，不进行微调或参数更新。模型架构和推理流程如图3所示。

重要的是，PathGen-CLIP未在Camelyon17上进行预训练或更新，从而确保我们实验中观察到的所有性能差异都完全源于输入数据的领域偏移，而非模型适配或重新训练。这避免了信息泄漏，也更准确地反映了现实中的监测场景，在这些场景中，部署数据在此之前对AI模型完全不可见。

在我们的设置中，Camelyon17的站点1、2和3被合并，用作模拟的部署参考数据集（表II），而不是使用PathGen-CLIP的实际训练数据。这种设计与实际条件相符：商业AI产品的原始训练集通常不可用或无法访问，因此会使用部分部署数据（模型在该数据上表现出预期性能）作为持续监测的参考基准。

C. 推理

在推理时，我们使用病理视觉语言模型（PathGen-CLIP）进行二元肿瘤分类，标签集为$\mathcal{C} = \{$**肿瘤**, **正常**$\}$。我们向模型提供两个文本提示：“一张肿瘤图像块的苏木精-伊红染色图像”和“一张正常图像块的苏木精-伊红染色图像”，作为视觉语言模型文本分支的类别描述。

推理时，模型接收两个输入：1）一个病理图像块；2）文本提示。图像编码器生成一个图像嵌入$\mathbf{z}$，文本编码器为每个类别 $c \in \mathcal{C}$生成一个嵌入$\mathbf{t}_c$ 。随后，模型计算图像嵌入与两个提示的嵌入之间的相似度（相关性）。这些相似度分数通过Softmax函数转换为概率。概率较高的类别将被选为最终预测。例如，如果某个病理图像块的嵌入与“一张肿瘤图像块的苏木精-伊红染色图像”的嵌入比与“一张正常图像块的苏木精-伊红染色图像”的嵌入更相似，则该图像块会被分类为肿瘤。

在数学上，我们在计算相似度之前，先对图像嵌入和文本嵌入进行 $\mathcal{L}_2$ 归一化，

$$\hat{z} = \frac{z}{\|z\|_2}, \qquad \hat{t}_c = \frac{t_c}{\|t_c\|_2},$$

然后使用CLIP logit缩放因子 α计算相似度分数，

$$s_c = \alpha\, \hat{z}^\top \hat{t}_c,$$

类别后验概率可通过以下方式获得：

$$p(y = c \mid x) = \frac{\exp(s_c)}{\sum_{c \in \mathcal{C}} \exp(s_c)}, \quad c \in \mathcal{C}.$$

对于每张病理图像，我们记录：1）512 维归一化图像嵌入；以及 2）针对两个类别（肿瘤与正常）的二维后验概率向量。这些存储的后验概率随后用于输出分析。

D. 数据偏移与性能下降分析

1) 实验设置：本节分析部署地点（分布内）与分布外站点（即 OOD-S1 和 OOD-S2）之间的数据偏移，并探究这种偏移与病理学视觉语言模型的肿瘤分类性能之间的关系。

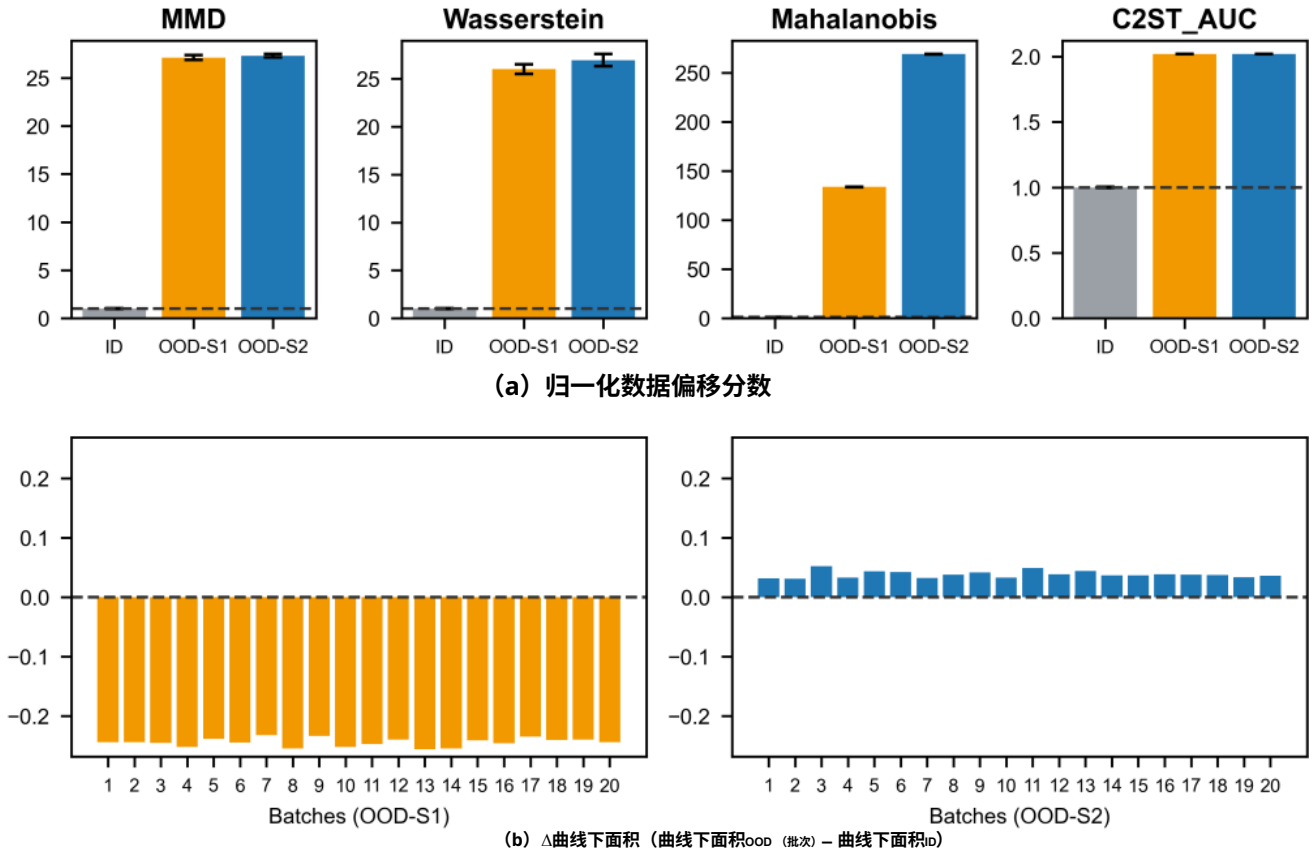


图 4。 最先进的病理视觉语言模型在分布外（OOD）站点上的输入数据偏移和性能下降。（a）相对于分布内（ID）数据集的基线，测量每个 OOD 站点的 20 个分布外子组（批次）的数据偏移分数（均值 ± 标准差）。（b）视觉语言模型在 20 个 OOD批次上的性能下降（曲线下面积下降）；基线是模型在完整分布内站点上的肿瘤分类曲线下面积。

偏移量化指标。为量化分布内站点与分布外站点之间的数据偏移，我们采用四种广泛使用的指标：*1)* 最大均值差异（*MMD*）；*2) Wasserstein*距离；*3)* 马氏距离；以及*4)* 分类器双样本检验（*C2ST*），其中使用逻辑分类器的曲线下面积作为偏移分数。

数据偏移基线。原始数据偏移分数通常难以解读，因为不同指标的尺度各异。为了提供有意义的参考，我们构建了一个分布内（ID）数据偏移基线。具体而言，我们从ID数据集中有放回地抽取20个分布内批次（每个批次包含5,000个样本，且肿瘤/正常比例均衡），并使用最大均值差异、Wasserstein距离、马氏距离和C2ST计算这些批次相对于完整ID数据集的偏移分数。各分布内批次的平均偏移分数用作每项指标的基线。

对于每个 OOD 批次，我们计算其相对于完整 ID数据集的偏移分数，并将其除以相应的基线值。因此，所有报告的偏移分数均表示 OOD 偏移是 ID 基线的多少倍。例如，如果基线最大均值差异为 5，而某个 OOD 批次的最大均值差异为 100，则报告的偏移分数为 20，这意味着 OOD 偏移是正常分布内变化的 20 倍。

OOD 站点的数据偏移计算。对于 OOD 站点 1，我们从整个 OOD 站点 1 中抽样，构建 20 个分布外子组（批次），每个子组包含 5,000 个样本（肿瘤/正常比例为 50:50）。对于每个批次 $i \in \{1, 2, \ldots, 20\}$ 在 OOD-S1 中，我们计算其相对于 ID 数据偏移分数基线的偏移分数（即最大均值差异、Wasserstein、马氏距离、C2ST）。对 OOD-S2 也采用相同的流程。

性能下降量化指标。我们首先计算分布内站点上的肿瘤分类性能（以曲线下面积计），记为 $\mathrm{AUC}_{\mathrm{ID}}$。对于 OOD-S1 中的批次$i \in \{1, 2, ..., 20\}$，视觉语言模型的肿瘤分类曲线下面积记为 AUC_i，则（在 OOD 批次 i 上的）性能下降定义为：

$$\Delta\mathrm{AUC}_i = \mathrm{AUC}_i - \mathrm{AUC}_{\mathrm{ID}}. \tag{1}$$

由于 OOD-S1 被划分为 20 个子组（批次），病理视觉语言模型会在每个批次上执行肿瘤分类。我们计算每个子组（批次）的性能下降，并考察其与数据偏移程度（即其与ID数据集的差异）的关系。对 OOD-S2 采用相同的流程。

2) 结果：图4 (a) 展示了 OOD-S1 和 OOD-S2 的数据偏移分数（20 个 OOD 批次上的均值±标准差）。所有 OOD 批次的偏移分数均显著高于基线。最大均值差异和 Wasserstein 距离的增幅较大（约为 $25\times$），而马氏距离的增幅更大（OOD-S1 约为 1.3×10^2，OOD-S2 约为 2.7×10^2）。C2ST分类AUC 相较基线也翻了一番，表明简单的分类器就能轻松区分 OOD 样本和 ID 样本。

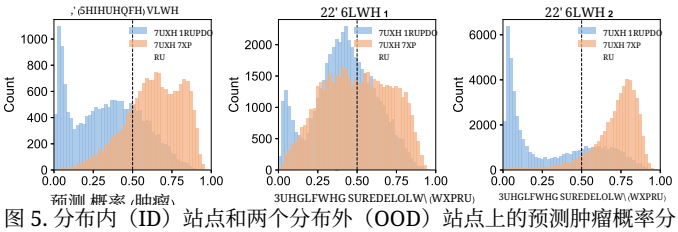


图 5. 分布内（ID）站点和两个分布外（OOD）站点上的预测肿瘤概率分布。每个面板展示真实正常（蓝色）和真实肿瘤（橙色）样本的预测肿瘤概率直方图。在 OOD 站点 1 上，预测结果集中在决策边界附近（**0.5**）。在 OOD 站点 2 上，预测结果的区分度更高。

面板 (b) 展示了病理视觉语言模型相对于其分布内性能的性能下降（AUCdrop）。对于 OOD-S1，该模型的曲线下面积在全部 20 个批次中均持续下降约 0.25。相比之下，对于 OOD-S2，变化幅度较小且略微为正（$0.03 \sim 0.05$），表明没有明显的性能下降。

3）分析：所有偏移检测方法均成功区分了分布外数据集与分布内数据集，表明这些方法能够有效捕捉病理图像中的外观变化。这证实，输入数据偏移检测能够提供有价值的诊断信息，并可作为数据分布变化和潜在性能波动的早期预警信号。

两个分布外站点均出现了较大的数据偏移；然而，只有 OOD-S1 出现了显著的性能下降。这一观察结果凸显了输入数据偏移是一种**无方向性**的综合指标：它能够捕捉数据外观的任何变化（例如扫描仪或染色差异），却无法表明这些变化是否会影响模型性能。

因此，大幅输入偏移并不一定会影响视觉语言模型的分类边界，因此可能属于无害偏移。相反，某些偏移会使输入样本移向决策边界或聚集在其附近，增加不确定性和误分类，进而导致显著的性能下降，属于有害偏移。因此，输入数据偏移检测是一种有用的诊断信号，表明“发生了变化”，但其本身并不能可靠地预测人工智能性能是否会出现有害的下降。

E. 基于输出的性能下降指标

1) 动机：从上述实验研究中，我们观察到输入数据偏移并不总会导致性能下降。尽管检测到较大的数据偏移有助于针对潜在的人工智能性能下降触发警报，但更直观的策略是直接监控并检测性能下降本身。我们的实验表明，基于输出的变化与性能变化的关系更为密切直接，这促使我们设计基于输出的性能下降指标。为说明这一动机，我们首先对病理视觉语言模型的输出进行可视化。

对于每张输入病理图像（图像块），病理视觉语言模型会输出两个概率：1）图像为肿瘤图像的概率，记为 p(肿瘤)；2）图像为正常（健康）图像的概率，记为 p(正常)。对于对于输入的**肿瘤** 图像，如果视觉语言模型表现良好，p(肿瘤)应较高（例如，0。8），而 p(正常)应较低（例如，0。2）。基于这一判断，我们将分布内站点和两个分布外站点中所有图像的 p(肿瘤)概率可视化。

图5 展示了 p(肿瘤) 在分布内站点和两个分布外站点上的分布，即，OOD-S1 和 OOD-S2。蓝色直方图表示 p(肿瘤)在正常（健康）病理图像上的分布，而橙色直方图表示 p(肿瘤)在肿瘤图像上的分布。从图5可以看出，在分布内站点，两个分布明显分离（正常样本集中在0附近，肿瘤样本集中在1附近）。这表明，对于真实正常图像，预测肿瘤概率较低；而对于真实肿瘤图像，预测肿瘤概率较高。这正是性能良好的病理人工智能模型所应有的表现。

然而，在 OOD-S1 上，两个分布有大量重叠。在这种情况下，视觉语言模型往往不会给出明确且两极分化的概率（例如，80% 肿瘤 vs.20% 正常），而是给出较为模糊的估计（例如，47%肿瘤 vs.53%正常）。因此，更多真实肿瘤图像被误分类为正常，反之亦然，导致性能显著下降。在 OOD-S2 上，情况则截然不同。尽管输入数据偏移非常显著（如上一节所量化并在图4中所示），真实正常图像和真实肿瘤图像的输出分布（关于 p(肿瘤)）仍然明显分离。因此，模型的性能（如图4所示）并未出现显著下降。

综上，这些模式表明，基于输出的指标对于检测医疗人工智能系统中的性能下降更加敏感，也能提供更直接的信息。

基于输出的性能退化指标：当病理视觉语言模型应用于分布外数据集 $\mathcal{S}$ 时，令 $p_i \in [0,1]$ 表示样本 i 属于正类（肿瘤）的预测概率。因此，模型会为 $\mathcal{S}$ 中的 n 个样本输出一组预测概率 $\{p_i\}_{i=1}^{n}$。为检测 $\mathcal{S}$ 上可能出现的性能下降，我们引入一种基于输出的指标，称为**基于置信度的退化指数（CDI）**。我们以两种简单的无标签形式实现 CDI：基于间隔的退化指数（*CDI M*）和基于熵的*CDI*（*CDI H*）。

基于间隔的退化指数（CDI M）。该指标衡量模型预测平均而言距离决策边界 $\mathcal{P}$ 有多近（即，$p_i > \mathcal{P}$ 表示肿瘤；否则为正常）。对于数据集 $\mathcal{S}$，我们定义

$$\mathrm{CDI_M}(\mathcal{S}) = 2\,\frac{1}{n}\sum_{i=1}^{n}\left|p_i - \mathcal{P}\right| \tag{2}$$

在我们的设置中，由于我们的任务是一个正类和负类大致均衡的二分类问题，因此将决策边界 $\mathcal{P}$ 设为 0.5。CDI M =0 对应于置信度非常低的预测（*i.e.*，$p_i \approx 0.5$）；而接近 1 的值则表示预测集中在 0 和 1 附近。因此，随着 CDI M 降低（CDI M↓），置信度也会降低，这可能表明性能下降（曲线下面积↓）。

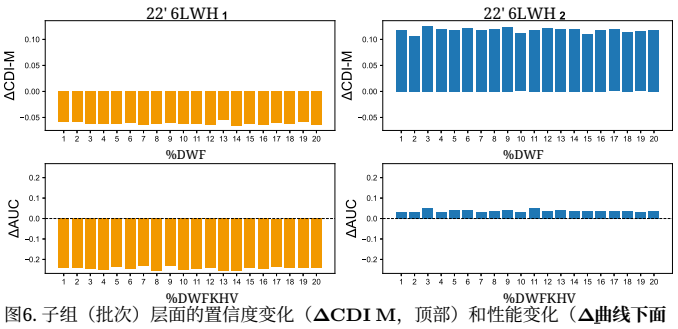


图6. 子组（批次）层面的置信度变化（**ΔCDI M**，顶部）和性能变化（**Δ曲线下面积**，底部），适用于 OOD-S1 和 OOD-S2。OOD-S1 显示出置信度崩塌，并伴有曲线下面积大幅下降；而 OOD-S2 的置信度有所提高，曲线下面积则保持稳定或略有改善。

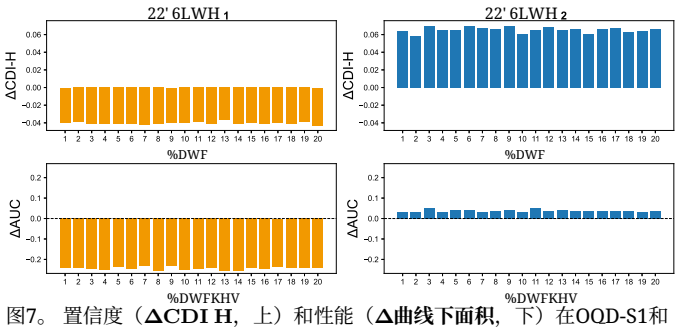


图7。 置信度（**ΔCDI H**，上）和性能（**Δ曲线下面积**，下）在OOD-S1和OOD-S2上的子组（批次）变化。OOD-S1表现出置信度崩塌，并伴随显著的AUC下降；而OOD-S2则表现出置信度提升，同时AUC略有改善。

为了检测医学人工智能模型在数据集 $\mathcal{S}_i$ 上相对于参考数据集 $\mathcal{S}_{\text{ref}}$的性能下降，我们计算

$$\Delta\text{CDI_M} = \text{CDI_M}(\mathcal{S}_i) - \text{CDI_M}(\mathcal{S}_{\text{ref}}). \tag{3}$$

负向 ΔCDI M 表示置信度崩塌，通常与性能下降相关；而**正向**值表示预测置信度更高，且没有性能下降。

基于熵的 CDI（CDI H）。该指标通过平均二元香农熵 [35]捕捉输出歧义。设预测概率 p的香农熵（对数底数设为 2）为

$$H(p) = -p\log p \; - \; (1-p)\log(1-p).$$

对于包含 n个样本的数据集 $\mathcal{S}$，我们定义

$$\text{CDI_H}(\mathcal{S}) = 1 - \frac{1}{n}\sum_{i=1}^{n} H(p_i). \tag{4}$$

CDI H的值反映了数据集 $\mathcal{S}$上的平均预测置信度。当预测完全确定时（即，$p = 0$ 或 $p = 1$），该值等于1（最大值）；随着预测逐渐接近决策边界（即，$p = 0.5$），该值会向0降低。因此，随着CDI H降低（CDI H ↓），置信度会下降，这可能表明性能出现潜在下降（AUC↓）。

为了检测医学人工智能模型在数据集 $\mathcal{S}_i$ 上相对于参考数据集 $\mathcal{S}_{\text{ref}}$的性能下降，我们计算

$$\Delta\text{CDI_H} = \text{CDI_H}(\mathcal{S}_i) - \text{CDI_H}(\mathcal{S}_{\text{ref}}). \tag{5}$$

CDI H为**负** Δ值反映置信度骤降，这通常与性能下降有关；而**正**值则表示置信度较高，且未出现性能下降。

3）结果：对于OOD-S1，我们通过从完整的OOD-S1中抽样，构建了20个分布外子组（批次），每个子组包含5,000个样本（肿瘤/正常比例为50:50）。对于OOD-S1中的每个批次 $i \in \{1, 2, \ldots, 20\}$，我们计算其ΔCDI M_i 和 ΔCDI H_i，以及AUC变化（即，ΔAUC_i）。对OOD-S2也采用相同的流程。

图 6 展示了 ΔCDI M （上排）以及相应的 Δ曲线下面积（下排），涵盖各 OOD 站点的全部 20 个子组（批次）。两个站点采用相同的纵轴范围，因此数值大小可以直接比较。

在**OOD-S1**上，每个批次均显示负的 ΔCDI M（≈ -0.05 至 -0.06），表明预测结果向决策边界伴随着不确定性增加和置信度丧失。同时， Δ曲线下面积始终为较大的负值（-0.23 至 -0.25），表明性能显著下降。

相比之下，在**OOD-S2**上，所有批次的 ΔCDI M均为 > 0（$\approx +0.09$ 至 $+0.12$），表明相较于ID，预测结果略微远离决策边界，反映出置信度提升。相应的 Δ曲线下面积值较小且略为正值（约为 $+0.03$ 至 $+0.06$），表明性能略有提升。

总体而言，置信度的变化与性能下降密切相关：置信度发生崩塌（ΔCDI M < 0）后，性能随之下降（AUC ↓）；而当置信度未发生崩塌（ΔCDI M > 0）时，性能保持稳定或略有提升。

图7 展示了 ΔCDI H以及 Δ曲线下面积在各分布外站点的情况。在**OOD-S1**上， ΔCDIH 始终为负，而曲线下面积大幅下降；相比之下，在**OOD-S2**上， ΔCDI H 为正，性能（曲线下面积）则保持稳定或略有提升。这些模式与CDIM的模式相似，证实该指标与性能变化密切相关。

4）讨论：在两个分布外站点，输出空间指标与性能变化呈现密切关系，这与概率直方图所示结果一致。当**基于置信度的退化指数**下降时，性能也会下降；当**基于置信度的退化指数**保持不变时，性能则保持稳定，甚至略有提升。这表明，输出空间指标能够提供有意义的预后信息：当预测趋于犹豫不决时，往往会随之出现性能退化；而歧义较低则对应着稳定或有所提升的性能。

这些发现表明，输入偏移检测与基于输出的指标各自承担着明确不同的作用。输入数据偏移检测告诉我们“发生了变化”，但并未揭示这种变化是否会损害病理视觉语言模型的性能。相比之下，基于输出的指标有助于判断模型是否正在经历实际退化。

在实践中，这两种方法互为补充。输入偏移检测适合用于发出早期预警信号，而基于输出的指标虽然处理略有延迟（相较于输入数据），却能更直接地提示性能下降。结合使用这些信号，可以更可靠地开展医学人工智能监测。

V. 结论

本研究考察了数据偏移情况下最先进的病理学视觉语言模型的性能下降，并探讨了输入级数据偏移检测和基于输出的指标在可靠性监测中的作用。我们表明，输入数据偏移检测能够识别分布变化并作为早期预警信号，提供有价值的诊断信息；但此类偏移并不总会导致性能下降。为补充输入层面监测，我们提出了一种轻量级、无标签的基于输出的置信度指标，可更直接地反映数据偏移情况下模型性能的下降。我们还开发了 DomainSAT，这是一款开源图形用户界面工具箱，可进行系统的偏移分析，并直观地展示特征空间变化。这些互补的输入层面和输出层面方法共同为真实世界部署环境中的病理学视觉语言模型可靠性监测提供了实用工具和见解。

致谢

本研究部分得到了美国国家医学图书馆的资助，项目编号为 1R01LM014239。

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh 等，“从自然语言监督中学习可迁移的视觉模型”，载于 国际机器学习会议，2021年，第8748–8763页。[2] Z. Wang, Z. Wu, D. Agarwal 和 J. Sun，“Medclip：从非配对医学图像与文本中进行对比学习”，载于自然语言处理实证方法会议论文集，2022年，第3876–3887页。[3] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon 和 J. Gao，“Llava-med：一天内训练用于生物医学的大语言与视觉助手”，*Advances in Neural Information Processing Systems*，第36卷，第28 541–28 564页，2023年。[4] J. Li, Y. Li, Z. Sun, E. Yu 等，“探索多模态大语言模型在经胸超声心动图（TTE）任务中的应用，以支持心血管决策”，《生物医学信息学杂志》，第171卷，第104930页，2025年。[5] S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane 和 S. Saria，“人工智能中的临床医生与数据集偏移”，《新英格兰医学杂志》，第385卷，第3期，第283–286页，2021年。[6] K. Stacke, G. Eilertsen, J. Unger 和 C. Lundstr¨om，“测量深度学习在组织病理学中的领域偏移”，《*IEEE*生物医学与健康信息学汇刊》，第25卷，第2期，第325–336页，2020年。[7] H. Guan 和 M. Liu，“医学图像分析的域适应：综述”，《*IEEE*生物医学工程汇刊》，第69卷，第3期，第1173–1185页，2022年。[8] ——，“Domainatm：医学数据分析的域适应工具箱”，《神经影像》，第268卷，第119863页，2023年。[9] H. Guan、D. Bates 和 L. Zhou，“保持医学人工智能的健康与可信：系统退化检测与纠正方法综述”，《*IEEE*生物医学工程汇刊》，2026年。[10] J. Li、D. Li、S. Savarese 和 S. Hoi，“Blip-2：利用冻结图像编码器和大语言模型引导语言-图像预训练”，载于国际机器学习会议，2023年，第19 730–19 742页。[11] H. Liu、C. Li、Q. Wu 和 Y. J. Lee，“视觉指令微调”，*Advances in Neural Information Processing Systems*，第36卷，第34 892–34 916页，2023年。[12] S. Zhang、Y. Xu、N. Usuyama、H. Xu 等，“Biomedclip：基于一千五百万科学图像-文本对预训练的多模态生物医学基础模型”，*arXiv*预印本 *arXiv:2303.00915*，2023年。[13] Y. Sun、Y. Zhang、Y. Si、C. Zhu 等，“Pathgen-1.6m：通过多智能体协作生成160万组病理图像-文本对”，载于第*13*届国际学习表征会议，2025年，第1–15页。[14] A. Kore、E. Abbasi Bavil、V. Subasri、M. Abdalla、B. Fine、E. Dolatabadi 和 M. Abdalla，“真实世界医学影像数据上的经验数据漂移检测实验”，Nature Communications，第15卷，第1期，第1–11页，2024年。[15] W. Wang、H. Li、Z. Ding、F. Nie、J. Chen、X. Dong 和 Z. Wang，“重新思考视觉域适应中的最大均值差异”，IEEE Transactions on Neural Networks and Learning Systems，第34卷，第1期，第264–277页，2021年。[16] S. Nietert、Z. Goldfeld、R. Sadhu 和 K. Kato，“切片瓦瑟斯坦距离的统计、鲁棒性与计算保证”，Advances in Neural Information Processing Systems，第35卷，第28 179–28 193页，2022年。[17] H. Ghorbani，“马氏距离及其在检测多变量离群值中的应用”，Facta Universitatis，《数学与信息学系列》，第583–595页，2019年。[18] B. Fuglede 和 F. Topsoe，“詹森-香农散度与希尔伯特空间嵌入”，载于《国际信息论研讨会》。IEEE，2004年，第31页。[19] J. M. Joyce，“库尔贝克-莱布勒散度”，载于《国际统计科学百科全书》。施普林格，2011年，第720–722页。[20] V. W. Berger 和 Y. Zhou，“柯尔莫哥洛夫–斯米尔诺夫检验：概述”，Wiley statsref: Statistics reference online，2014年。[21] R. S. M. de Barros、J. I. G. Hidalgo 和 D. R. de Lima Cabral，“威尔科克森秩和检验漂移检测器”，Neurocomputing，第275卷，第1954–1963页，2018年。[22] D. A. Darling，“柯尔莫哥洛夫–斯米尔诺夫检验与克拉默-冯·米塞斯检验”，The Annals of Mathematical Statistics，第823–838页，1957年。[23] J. Merkow、A. Soin、J. Long、J. P. Cohen 等，“CheXstray：用于医学影像人工智能的实时多模态监测工作流”，载于《医学图像计算与计算机辅助干预国际会议》。施普林格，2023年，第326–336页。[24] S. Dreiseitl，“医学数据集上协变量偏移检测方法的比较”，载于《计算机辅助系统理论国际会议》。施普林格，2022年，第487–495页。[25] D. Lopez-Paz 和 M. Oquab，“重新审视分类器双样本检验”，载于《国际学习表征会议》，2017年。[26] R. S. M. de Barros、J. I. G. Hidalgo 和 D. R. de Lima Cabral，“关于使用随机森林进行双样本检验”，Computational Statistics & Data Analysis，第170卷，第107435页，2022年。[27] J. B´obeda、M. J. Garc´ıa-Gonz´alez、L. V. P´erez-Herrera 和 K. L´opez-Linares，“使用卷积自编码器进行无监督数据漂移检测：乳腺癌影像场景”，载于《KES医学与医疗保健创新国际会议》。施普林格，2023年，第345–354页。[28] D. Lopez-Paz 和 M. Oquab，“重新审视分类器双样本检验”，载于《国际学习表征会议》，2017年，第1–15页。[29] M. Kirchler、S. Khorasani、M. Kloft 和 C. Lippert，“使用深度学习进行双样本检验”，载于《人工智能与统计国际会议》，2020年，第1387–1398页。[30] T. Pandeva、T. Bakker、C. A. Naesseth 和 P. Forr´e，“评估分类器双样本检验”，Transactions on Machine Learning Research，第1–15页，2024年。[31] L. M. Koch、C. F. Baumgartner 和 P. Berens，“用于医学人工智能上市后监测的分布偏移检测：一项回顾性模拟研究”，npj Digital Medicine，第7卷，第1期，第1–11页，2024年。[32] P. Bandi、O. Geessink、Q. Manson、M. Van Dijk 等，“从单个转移灶检测到患者层面的淋巴结状态分类：CAMELYON17挑战赛”，IEEE Transactions on Medical Imaging，第38卷，第2期，第550–560页，2018年。[33] P. W. Koh、S. Sagawa、H. Marklund、S. M. Xie 等，“WILDS：真实环境中分布偏移的基准”，载于《国际机器学习会议》，2021年，第5637–5664页。[34] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh 等，“从自然语言监督中学习可迁移的视觉模型”，载于《国际机器学习会议》，2021年，第8748–8763页。[35] J. Lin，“基于香农熵的散度度量”，IEEE Transactions on Information Theory，第37卷，第1期，第145–151页，2002年。