arXiv:2611.32899v1 [cs.CV] 24 Nov 2026

通过文件头导致的标签泄漏：公开乳腺摄影数据集中的几何与批次混杂因素，以及预训练验收门槛

Achraf Toufah ORCID 0009-0006-6802-3905
Qanary Inc.
Rana Saad ORCID 0009-0005-5806-2604
Qanary Inc.
Moulay Youssef El Hafidi MYELHAFIDI@QANARY.NET
Qanary Inc.（通讯作者；*ORCID 0000-0001-7413-3293*）

摘要

如果一个数据集的诊断标签可以从非诊断信息中推断出来，就会产生很高、可复现但没有信息价值的模型性能。我们对六个公开乳腺摄影数据集实施了预训练控制，只考察图像几何、全局纹理、子集合标识符或一个协变量，并以曲线下面积 0.60 作为停止阈值。在 KAU-BCMD 中，仅凭几何信息就能在本地副本中以曲线下面积 1.0000 预测 BI-RADS 类别，在公开 Kaggle 存档中的均值则为 0.979；调整尺寸后，仅凭纹理信息的曲线下面积为 0.847。在 CMMD 中，子集合前缀以 0.7753 的曲线下面积预测恶性肿瘤；受限子集合通过了几何（0.5000）和纹理（0.5810）控制，但未通过事后添加的标识符顺序控制（测试集曲线下面积为 0.9926）。CBIS-DDSM 在几何控制上超过停止阈值（20 个划分的合并均值为 0.654），这是一个事后选定的判定统计量；其官方肿块与钙化划分则未超过阈值。M-SYNTH、INbreast 和部分 VinDr-Mammo 副本通过了实际执行的少数几项控制。我们发布了代码和派生表。推理路径中不包含量子组件。

关键词：乳腺摄影、数据集质量、标签泄漏、批次效应、捷径学习、数据溯源、验收测试

1 引言

只有当标签无法通过其他方式获取时，关于医学图像的机器学习结果才是在描述解剖结构。如果还能从不包含诊断信息的因素中推断出标签，例如文件宽度、导出步骤留下的清晰度，或记录在列表中的位置，那么在该数据集上训练的每个模型都可能取得不错的分数，且这些分数可以复现，但这种复现并不能说明疾病情况。这一普遍现象被称为深度神经网络中的捷径学习 [1] ，也被称为“聪明的汉斯”式预测 [2], ；而在更早之前，计算机视觉领域就已对数据集偏差进行过研究[3]。

医学影像领域已有具体案例。一项横断面研究发现，用于检测胸部X线片中肺炎的深度学习模型，其泛化性能表现不一 [4]。在骨盆X线片中，一个深度学习模型利用 con- 预测髋部骨折患者及医疗变量：仅凭图像即可预测扫描仪型号，曲线下面积达到 1.00；而在一个

©2026 Achraf Toufah、Rana Saad 和 Moulay Youssef El Hafidi。

患者及医院流程变量之间的骨折风险均衡的测试集上，骨折预测则降至机会水平 [5]。研究人员探讨了皮肤镜图像中的手术皮肤标记与用于黑色素瘤识别的卷积网络诊断性能之间的关联 [6]。据报道，隐藏分层会导致医学影像机器学习出现具有临床意义的失效 [7]。针对放射影像 COVID-19 检测，有研究报告模型选择了捷径而非信号 [8], ；针对基于 X 射线图像的 COVID-19 自动检测方法，也有研究对其进行了严格评估 [9]。相关研究综述了医学影像机器学习中的此类方法学缺陷，并提出了建议 [10]。

相邻的两类文献用不同术语描述了同一缺陷。泄漏已被认为是造成基于机器学习的科学研究可复现性危机的因素之一 [11]。在高通量生物学中，测量结果会受到实验室条件、试剂批次和人员的影响，因而产生批次效应；当批次与关注的结果相关时，这会成为一个重大问题 [12]；在多中心神经影像数据集中，也已发现并纠正了跨站点偏倚 [13]。具体到乳腺摄影，Samala 等人模拟了乳腺肿块分类中的特征泄漏，并显示相较于独立测试集，验证性能可能被大幅高估 [14]；Elme-jgari 等人在 CBIS-DDSM 上量化了患者级数据泄漏，其中包括一种方案：在交叉验证前，将官方训练集和测试集划分合并 [15]；Liao 和 Aagaard 发布了一个开放代码库，旨在提高 CBIS-DDSM 深度学习诊断的透明度 [16]；Mabrok 比较了 CMMD 上按侧别评估与患者级评估 [17]；Zufiria 等人分析了乳腺摄影数据集在深度学习模型开发中可能存在的偏倚 [18]；Garrucho 等人则在一项多中心研究中考察了肿块检测的领域泛化 [19]。

报告实践主要通过文档和检查清单作出回应：数据集数据表 [20], 、科学数据管理与数据 stewardship 的 FAIR指导原则 [21], ，以及旨在为作者和审稿人提供指导的医学影像人工智能检查清单 [22]。大型公共医学图像数据集在发布后得到过研究 [23], ；一项依据 FAIR原则对乳腺摄影领域八个机器学习数据集进行的评述发现，这些数据集差异显著，既有数字采集图像，也有扫描胶片，而且在许可、访问便利性、标注可靠性和文件格式方面各不相同 [24]。文档描述的是数据集本身；仅凭文档并不能检验能否从文件中预测标签。

本文考察六个数据集：阿卜杜勒-阿齐兹国王大学乳腺癌乳腺X线影像数据集（KAU-BCMD） [25]；中国乳腺摄影数据库（CMMD）[26, 27], ，由癌症影像档案库 [28]发布；CBIS-DDSM，即乳腺筛查数字数据库的精选子集 [29]；INbreast [30]；VinDr-Mammo [31]；以及模拟的 M-SYNTH 队列 [32]。这类数据集也会被合并：一项基准数据汇编整合了来自六个来源的图像，其中包括 DDSM、INbreast、KAU-BCMD 和 CMMD [33]。当与标签相关的批次会产生最大影响时，数据汇集便是关键所在。

我们报告了对这六个数据集进行少量文件级控制后得到的结果，并考察了多种配置（KAU-BCMD 作为本地副本和公开 Kaggle 存档；CMMD 合并数据以及限定为一个子集合的数据）。三个数据集至少未通过一项停止角色控制。在 KAU-BCMD 中，BI-RADS 分类以目录归属的形式编码；仅使用图像宽度、高度、百万像素数和宽高比的分类器，便能将 BI-RADS 4/5级与 BI-RADS 1

区分开来：在本地副本上的曲线下面积为 1.0000，在公开 Kaggle 存档上 20 次按患者分组的划分中的均值为 0.979；两个副本中，两个类别之间都没有共享的图像尺寸；调整尺寸至相同的 224×224 后，纹理控制在公共档案上的结果仍为 0.847。在 CMMD 中，仅凭患者标识符的子集合前缀即可预测恶性肿瘤，曲线下面积为 0.7753。在 CBIS-DDSM 中，仅凭几何信息就与病理学相关：合并队列的平均曲线下面积为 0.654，钙化队列为 0.688，肿块队列为 0.635；这一结论依据的是看到结果后才选定的重复划分均值，而在单次官方划分中，肿块队列和钙化队列均低于停止阈值。三个数据集通过了对其进行的所有控制，但实际运行的控制很少：M-SYNTH（两项控制）、INbreast（仅几何，0.486）以及 VinDr-Mammo 可用的 59.8% 子集（仅几何，0.573）。

仅使用 CMMD 第一个子集合时，它不属于这两组中的任何一组。它通过了注册的几何和纹理控制，随后未能通过我们事后添加的一项控制：仅凭患者标识符中的序列号，就能以 0.9926 的测试曲线下面积预测恶性肿瘤。我们将此报告为本文自身的示例，说明通过注册控制并不等于获得证书。

相较于上文引用的研究，本文新增了测量、一个流程和一个发布版本，而没有提出新方法。测量内容是针对已命名公共数据集及其副本计算的对照组AUC，其中也包括结果平平无奇的数据集。流程是在训练目标模型之前，以书面形式确定一小组信息贫乏的控制变量及一个阈值，并记录在看到结果后作出了哪些分析选择。发布内容包括代码、附有输出的笔记本以及派生表。我们不主张优先权。我们使用方法部分列出的七个检索词查询了CrossRef书目索引，并检查了每个检索词返回的前六条记录；其中包括一篇关于CBIS-DDSM 患者级数据泄漏的报告 [15]和一项关于 CMMD 的评估研究 [17], ，但根据标题判断，这些记录均未报告 KAU-BCMD 或 CBIS-DDSM 中几何与标签之间的关联，也未报告 CMMD 标识符与恶性肿瘤之间的关联。这不是系统性综述，也未检查每项查询返回的前六条以外的记录。我们通过同日进行的网络搜索找到两篇预印本，并阅读了其摘要；它们与本文密切相关。Hajishafiezahramini等人 [34] 发现，在对三个乳腺摄影数据源（包括 CBIS-DDSM 和 CMMD）进行相同预处理后，分类器几乎可以完美地区分数据集来源；此外，加入来自外部的阳性病例会降低筛查性能。Zafari等人 [35]介绍了一种协调流程，并量化了 CBIS-DDSM、CMMD 的一个衍生版本和 VinDr-Mammo 之间的分布差异。两项研究都关注数据集之间的差异；本文报告的控制则考察能否从单个数据集中的非诊断信息恢复标签。

这项工作源于在临床乳腺图像上开展量子与经典机器学习模型对比基准测试的筹备工作，而这一上下文也解释了为何要进行对照实验。如果不在本地 KAU-BCMD 副本上运行这些对照实验，我们预计计划中的基准测试的每个组别，无论量子组还是经典组，得分都会接近曲线下面积 1.0，而量子组可能会被报告为约 0.99；这样的结果将具有可复现性、内部一致性，却没有信息价值。最近提出的量子优势框架 [36], 是一项提案，并非标准，将输出结果经过验证的正确性列为其两项标准中的第一项；如果数据集的标签可以从其文件头中读取，那么在进行任何计算之前，该数据集就已经对所有组别同样地不满足这一要求

表1：**对照组。“停止阈值”对照组在超过阈值时终止研究；协变量控制仅作为上下文报告。**

对照组	输入	Role	阈值	状态
几何	图像宽度，高度，百万像素，宽高比；无像素内容	stop	约 0.50 预期值；0.60 停止阈值研究	此前已注册 CMMD 运行
纹理	拉普拉斯算子方差，梯度能量，强度标准差，对比度范围，经过统一的调整大小；没有	stop	约为 0.50 预期值；0.60 则停止该研究	预先注册 CMMD 运行
批次 / 站点	几何 the 采集批次或子集合标识符本身	stop	约为 0.50 预期值；0.60 停止该研究	在……之后添加 CMMD纹理 纹理控制失败
单一协变量	单个人口统计变量本身，例如年龄	仅作为上下文，绝不作为停止阈值	none	在……之前注册 CMMD 运行
标识符顺序	数字（序列）部分患者标识符单独	stop	约 0.50 预期值为 0.60；停止该研究	事后添加，在……之后注册控制已通过 CMMD D1

了。本文没有报告量子优势、加速或量子增强性能。截至目前，更广泛项目中的每次量子运行都是经典态矢量模拟。推理路径中不包含量子组件。

---

2 方法

2.1 对照与停止规则

每个对照组都是一个分类器，其训练输入不应包含任何诊断信息，并以受试者工作特征曲线下面积（曲线下面积） [37] 相对于诊断标签进行评分。预期对照组的得分约为 0.50。

每个对照组使用的分类器都是 scikit-learn 的 HistGradientBoostingClassifier，采用默认超参数和随机_状态=0。我们刻意不进行调优：如果对照组经过调优，就可能有人质疑门控是经过优化以便失败的。

在对 CMMD 运行任何对照组之前，已书面确定 0.60 为停止阈值。注册的表述指出，几何和纹理控制在冻结测试集上的曲线下面积应约为 0.50；如果

其中任一项“超过 0.60”，则停止基准测试并调查数据集。后续笔记本将此规则实现为：当曲线下面积均值大于或等于 0.60 时判定失败。这两种表述仅在边界值上有所不同。本文报告的数值均不处于边界，因此这一差异不会影响任何结论。

该阈值是一项筛查规则，而非统计检验。它不会为结论附加p值或错误率，也不被视为区分可接受信号与问题泄漏的通用界限：低于阈值的对照组并不能证明数据集不存在捷径，而高于阈值的对照组也无法衡量下游模型会在多大程度上利用这些捷径。该阈值的作用是用于实际决策，判断是继续训练，还是先调查数据集。三项性质支持这种用法。第一，这些控制变量不包含诊断信息，因此其预期曲线下面积为0.50；在本研究中，对于结果接近机会水平的控制变量，0.60远远超出了划分间的变异范围：20次重复划分中最大的标准差为0.0217（INbreast几何，样本量最小的数据集），这意味着0.60比0.50高出4.6个标准差；而四个均值与机会水平相差不超过0.05的控制变量，其任何一次划分的结果都未超过0.5570。第二，在方差相等的双正态模型下，曲线下面积为0.60对应的标准化均值差为0.36，属于小效应，因此该规则标记的是规模虽小、但大于此处所见抽样变异的标签信息。第三，所有结论均在0.55、0.65、0.70和0.75这些阈值下重新评分；哪些结论取决于阈值选择、哪些不取决于阈值选择，将在第5节说明。

设计中有两项特征值得注意。几何控制不会解码任何像素；它会从每个文件头中读取四个整数。报告了协变量控制的结果，但从未将其用作停止阈值，因为人口统计变量与结果之间的关联可能是临床人群本身的特性，而非数据集的缺陷。

2.2 注册状态和偏离

文中所指的注册始终是内部制定并注明日期的书面方案。该方案并未存入公共注册平台。该方案及其注明日期的偏差日志均属于发布版本的一部分（数据可用性）。

几何、纹理和单协变量控制，以及 0.60 阈值，均在训练任何模型之前针对 CMMD 完成注册。针对本地副本进行的 KAU-BCMD 控制未注册：这些控制是在基准测试设置期间作为临时诊断运行的，门控则是在发现相关结果后才据此写入方案。因此，本地副本上的 KAU-BCMD 结果属于探索性结果。发现 CMMD 混杂因素后，批次/站点控制被提升为停止阈值。标识符顺序控制则是在之后才加入的（第 2.6 节）。M-SYNTH 控制依据另一项研究的方案运行。

另外两项分析在执行前完成了注册，并在后续一轮分析中实施，但存在以下偏离。下文中，“初始”始终指第一轮对照组分析（表2），“后续”指第二轮分析，其结果基于重复数据划分进行评分（表3）。

1. 在机构账户下，为公共 KAU-BCMD 分发版注册了新的下载，并登记了两条发布途径。实际只通过其中一条途径运行：由团队成员下载并保存在云端副本中的 Kaggle 托管档案；评估时采用了 20 次重复的按患者分组的划分，而非单次划分。

表2：**初始阶段分析的数据集及副本。**

~	**KAU-BCMD** [25], **带标签的PNG文件夹（本地副本）**	**CMMD** [26, 27]	**M-SYNTH** [32]
标签	BI-RADS 分类, 编码为目录 归属；无病理学 副本中的金标准	良性/恶性依据 活检，记录于单独的 元数据文件	病灶存在/不存在依据 构造
n	1,991 张PNG图像 · 501 · 1,000例病例-侧别单元	1,775名患者 · 1,872 乳房 · 5,202张图像	1,200例病例
类别平衡	60个BI-RADS 4–5级乳房 与 938 例 BI-RADS 1 相比	1,316 例恶性 / 556 例 良性乳房（70.3% 恶性）	按构造平衡
格式	8位 PNG ，不统一的 图像尺寸	MONOCHROME2 ，2294 × 1914 （4.39 兆像素）； 5,200 张 8 位图像和 2 张 16	模拟 ，256 × 256 补丁
许可	知识共享署名 4.0 国际许可协议	知识共享署名 4.0 国际许可协议	CC0
评分	一次按患者分组的 五折交叉验证 （分层分组K折）	在固定测试集划分上一次	在固定测试集划分上一次

2. 为另外五个公共数据集注册了门控：CBIS-DDSM、VinDr-Mammo、INbreast、RSNA 筛查乳腺摄影挑战赛数据集 [38] 和 BreastM-NIST [39], ，每个数据集均设置了四项控制变量，并在冻结划分上进行位深检查。五个数据集中实际运行了三个（CBIS-DDSM、INbreast、VinDr-Mammo），且仅采用几何控制；每个数据集均采用 20 次重复的患者分组五折划分进行评估；对于存在官方划分的数据集，同时报告其单次划分的曲线下面积和均值。

0.60 的阈值和 20 次数据划分的设计在后续运行前就已固定。将阈值应用于 20 次数据划分的均值，而非单次数据划分，是在看到结果后才作出的决定。这一点对 CBIS-DDSM（第 3.5 节）很重要。

2.3 分析的数据集与副本

KAU-BCMD 的数据集论文指出，该数据集由三位放射科医师进行标注和审核 [25]。CMMD 的数据集论文介绍了两个分支：一个包含活检确诊的良性或恶性标签，另一个包含已知的分子亚型 [26]；因此，CMMD 的两部分结构是其作者记载的内容，并非本研究的发现。该论文报告的乳腺X线影像数量少于我们分析的存档清单中的图像文件数量。我们没有核对这两个数量，本文中所有 CMMD 数量均取自清单和标签文件。

患者分组方式如下：KAU-BCMD 按文件名中的年份和病例标识符分组，并通过仅按病例编号分组进行敏感性分析；CBIS-DDSM 按两个队列中的患者标识符分组；INbreast 按文件名中的匿名患者标识符分组；VinDr-Mammo 则以研究标识符作为患者的代理标识符进行分组。

不同数据集的评分方案并不相同，我们不会将其视为等同。由于从未提供冻结划分，本地的 KAU-BCMD 副本采用了交叉验证

表3：**在后续轮次中采用几何控制进行分析的数据集及副本。**

~	**KAU-BCMD [25], 公开 Kaggle 存档**	**CBIS-DDSM [29]**	**INbreast [30]**	**VinDr-Mammo [31]**
来源	托管于 Kaggle 的 档案 king-abdulaziz-university-mammogram-dataset.zip （606,558,154 字节, 成员日期为 2022-03-01）	癌症影像 档案库 [28], 官方 下载客户端	原始 分发, “INbreast 发布版本 1.0”	PhysioNet; 下载 不完整
标签	BI-RADS文件夹；1 与4/5相对	来自.....的病理学 官方 病例描述 文件；恶性 与良性病例相比 （包括良性病例 未召回）	BI-RADS 1/2级 对比4/5 （评估，而非 病理学）	乳腺级别 BI-RADS 1/2级 对比4/5 （评估，而非 病理学）
分析数量	1,991 张 JPEG图像 · 501位患者 （BI-RADS 1/4/5级）	2,853 张全视野图像 图像 · 1,460 患者	379 张图像 · 103 患者	11,404 张图像 · 2,993 项检查
格式	8位 JPEG	16位 DICOM （数字化胶片）	16位 DICOM	12位和16位 DICOM
许可	知识共享署名 4.0 国际许可协 议	知识共享署名 3.0 以及 档案的数据使用 政策	访问协议 与提供方签订； 该协议并未 存档于 作者处	PhysioNet 受限健康 《数据》许可证 1.5.0 （凭证 访问）
覆盖率	完整（2,378 张中的 档案中的 2,378 张图像 ）	284 个官方项目中的 37 个 钙化测试 乳房视图 存在；所有其他 官方数据集至少 99.7%	完整（410 张，共 410 张图像）	20,000 个中的 11,965 个 图像（59.8%）

为其创建：数据集在划分前便已停止处理。CMMD 和 M-SYNTH 均在一个冻结划分上评分一次。可用的控制变量也有所不同，因为记录的元数据不同：M-SYNTH 没有患者标识符前缀，也没有年龄，因此无法对其进行批次和协变量控制；其纹理控制采用的是包含 15 个特征的整幅图像物理特征库，而非临床数据集所用的四特征特征库。

2.4 KAU-BCMD 的预处理修正方法

依次采用了三种标准修正方法，并重新评估对照组：将每幅图像调整到统一的 224×224；先将短轴统一为 467 像素，再调整尺寸；以及将范围限制为低于 0.5 百万像素的分辨率匹配子集。在两种调整尺寸的条件下，几何特征按构造为常量，因此仅使用纹理作为控制变量。

初步分析列出了四种纹理统计量，但未对其进行数值定义。在针对公共档案的后续重新运行中，这些统计量被明确规定如下：8位灰度；双线性重采样；拉普拉斯方差定义为 4 邻接离散拉普拉斯算子的方差；梯度能量定义为通过中心差分计算的 $dx^2 + dy^2$ 的均值；强度标准差定义为像素标准差；对比度范围定义为第 99 百分位数减去第 1 百分位数。因此，无法证明本地与公共纹理值是按完全相同的定义计算的。

2.5 CMMD冻结划分

CMMD划分依据患者最严重的发现进行分层，并按患者分组，随机种子为20261020；在训练任何模型之前，这些划分已写入磁盘。分析单位为乳房，因为97名拥有两侧已标注乳房的患者中，有30名患者两侧的分类不同。在合并数据集上纹理控制未通过后，方案修订为仅使用第一个子集合（D1），并在训练任何模型之前重新冻结了数据划分。发布的划分文件即为这些D1划分：605个训练乳房、171个验证乳房和331个测试乳房；测试集划分包含168个恶性乳房和163个良性乳房。

2.6 标识符顺序控制（事后）

CMMD患者标识符的格式为D1-0001。在注册门控通过CMMD D1后，将标识符的数字部分作为同一未调优分类器的唯一特征。该分类器在冻结的D1验证集划分和测试集划分上进行了评分，并在1,107个D1乳房的20组重复患者分组五折划分上进行了评分。该控制并非四项注册控制之一。它属于探索性分析，在文中出现时均标注为事后。

随后，对 CBIS-DDSM 也应用了同样的控制方法，其患者标识符的形式为 `P_00016`。标签取自档案 [28]，的四个公开病例描述文件，未读取任何图像。分析单位为图像（队列内患者、侧别和视图）；若图像上存在任何恶性病灶，则标记为阳性，与几何控制中的做法相同。对每个队列、两个队列合并后的数据（跨队列按患者分组），以及每个队列的官方训练集和测试文件，均进行了 20 次重复的患者分组五折划分评分。未对 INbreast 或 VinDr-Mammo 运行此项分析，因为它们的副本存放在

此次补充无法访问的云存储中；也未对 KAU-BCMD 运行，因为它在几何控制中已不合格。

2.7 统计

初始控制变量各报告一个曲线下面积，使用随机_状态=0运行一次。后续控制变量报告的是在 20 次重复的患者分组五折划分中，折外 AUC 的均值、标准差（SD）、最小值和最大值（使用带随机打乱的 StratifiedGroupKFold，划分随机种子为 0–19；模型固定为随机_状态=0，因为对于给定的训练集，分类器具有确定性），并根据均值作出结论。所有控制变量均未报告 p值，也未进行多重性校正：门控是针对绝对阈值进行筛查，而不是检验零假设；进行多重性校正反而会降低安全筛查的严格程度。初始控制变量没有可用的离散度数据。

后续分析使用 Python 3.13.5、scikit-learn 1.9.1、pandas 3.1.0、numpy 2.5.4、pydicom 3.0.2 和 Pillow 13.0.0 运行。初始分析所用的软件版本未作记录。

2.8 无量子组件

本研究中没有构建或评估任何电路，无论是模拟电路还是其他类型的电路。门控是一个基于信息贫乏输入的梯度提升分类器。

2.9 文献检索

引言中关于早期报告的陈述，基于本文撰写期间对CrossRef书目索引进行的查询；对以下每个检索词组，均检查了前六条记录：“KAU-BCMD乳腺X线图像数据集分类深度学习”；“CMMD中国乳腺摄影数据库数据泄漏批次子集合”；“CBIS-DDSM数据泄漏患者重叠训练测试集划分”；“捷径学习乳腺摄影伪相关图像尺寸数据集偏差”；“公开乳腺X线摄影数据集重复图像INbreast CBIS-DDSM泄漏”；“标签泄漏文件元数据DICOM头医学影像数据集”；以及“图像分辨率混杂因素深度学习医学图像数据集类别标签图像尺寸”。

---

3 结果

3.1 概述

表4列出了所有对照结果。图 2显示了相同的行。

三个数据集未通过至少一项停止角色控制：KAU-BCMD、CMMD 和 CBIS-DDSM。三个数据集通过了对其执行的所有控制：M-SYNTH（两项控制）、INbreast（仅几何控制）以及 VinDr-Mammo 的可用子集（仅几何控制）。仅限于 D1 的 CMMD 通过了两项已注册的停止控制，但未通过事后添加的标识符顺序控制，因此不计入通过的数据集。CBIS-DDSM 通过了标识符顺序

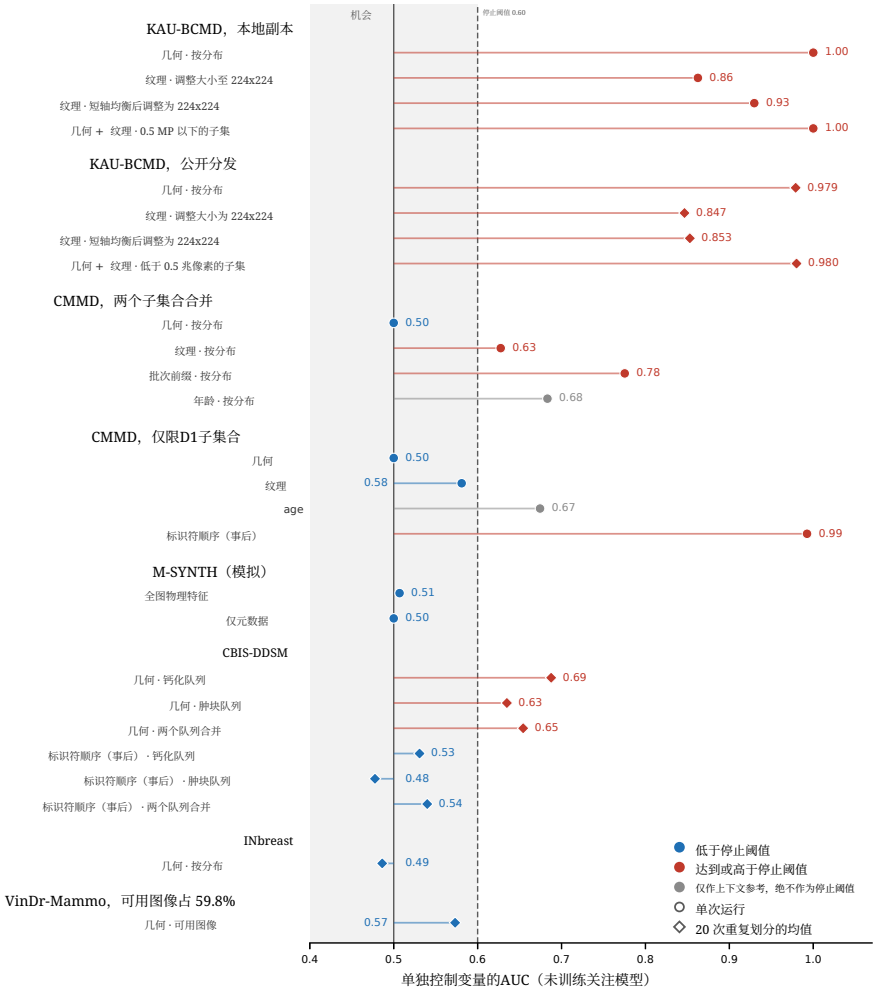


图1：**对六个公开乳腺摄影数据集进行预训练控制**。每个标记表示单独一种控制与诊断标签之间的曲线下面积（横轴，“单独控制变量的AUC（未训练感兴趣模型）”）；虚线表示 0.60 的停止阈值。标记颜色表示结论：蓝色，低于停止阈值；红色，达到或高于停止阈值；灰色，上下文（单协变量控制，仅作报告，不会使研究停止）。圆形表示单次运行，菱形表示 20 次重复划分的均值；数值标注在各标记旁边。自上而下的区块：KAU-BCMD 本地副本（几何；调整尺寸后的纹理；短轴均衡化后的纹理；低于 0.5 兆像素子集上的几何 + 纹理）；KAU-BCMD 公开 Kaggle 存档（相同的四行）；CMMD 合并（几何；纹理；批次前缀；年龄）；CMMD D1（几何；纹理；标识符顺序；年龄）；M-SYNTH（全图物理特征；仅元数据）；CBIS-DDSM（几何：钙化、肿块、合并；标识符顺序：钙化、肿块、合并）；INbreast（几何）；VinDr-Mammo 可用子集（几何）。单次运行：KAU-BCMD 本地副本各行（一次按患者分组的五折交叉验证），以及 CMMD 和 M-SYNTH 各行（在固定测试集划分上对一次拟合结果进行评分；批次前缀的数值根据标签文件计算）。20 次重复的患者分组五折划分均值：KAU-BCMD 公开档案各行，以及 CBIS-DDSM、INbreast 和 VinDr-Mammo 各行。标识符顺序各行均为事后添加；对于 CMMD D1，该控制是在该队列通过已注册控制后添加的，属于探索性分析。CBIS-DDSM 的结论依据 20 次划分均值，这是在查看结果后选定的统计量；在单次官方划分中，肿块和钙化

表4：**对每个对照组、每个数据集和每种配置均进行了测量**。“单次运行”是指在随机_状态=0下拟合一次；“20次划分均值”是指在20次重复的患者分组五折划分中，折外曲线下面积的均值，括号内为标准差和范围。

数据集及配置	对照组	AUC	评分	结论
KAU-BCMD，本地副本，按其保存的状态	几何	1.0000	单次运行（一按患者分组五折交叉验证）	fail
KAU-BCMD，本地副本，调整大小后至 224 × 224	纹理	0.8625	单次运行（如上文）	fail
KAU-BCMD，本地副本，短轴均衡至467 像素，然后 224 × 224	纹理	0.9297	单次运行（如上文所述）	fail
KAU-BCMD，本地副本，子集低于 0.5 兆像素	几何 + 纹理	1.0000	单次运行（如上文所述）	fail
KAU-BCMD，公开Kaggle档案，如分布式	几何	0.9790（标准差 0.0082；0.9712–0.9995）	20次划分均值	fail
KAU-BCMD，公开Kaggle档案，调整大小至 224 × 224	纹理	0.8465（标准差 0.0110；0.8233–0.8641）	20次划分均值	fail
KAU-BCMD，公开Kaggle档案，短轴均衡至 467 像素，然后 224 × 224	纹理	0.8529（标准差 0.0097；0.8342–0.8743）	20次划分均值	fail
KAU-BCMD，公开 Kaggle档案，子集低于 0.5 兆像素	几何 + 纹理	0.9800（标准差 0.0112；0.9649–1.0000）	20次划分均值	fail
CMMD，合并 D1 +D2	几何	0.5000	单次运行，冻结测试集划分	pass
CMMD，合并 D1 +D2	纹理	0.6275	单次运行，冻结测试集划分	fail
CMMD，合并 D1 +D2	批次前缀（D1 仅与 D2 对比）	0.7753	单一计算来自标签文件	fail
CMMD，合并 D1 +D2	仅年龄	0.6832	单次运行，冻结测试集划分	上下文，而非 stop
仅 CMMD D1，数据划分重新冻结	几何	0.5000	单次运行，已冻结测试集划分	pass
仅限 CMMD D1，数据划分重新冻结	纹理	0.5810	单次运行，已冻结测试集划分	pass
仅限 CMMD D1，重新冻结数据划分	标识符顺序 [事后]	0.9926	单次运行，冻结测试集划分（20次划分均值 0.9871；表 8）	失败（添加对照组事后）
仅 CMMD D1，重新冻结的数据划分	仅年龄	0.6744	单次运行，冻结测试集划分	上下文，而非 stop
M-SYNTH，冻结测试集划分，n = 352	整幅图像物理特征	0.5070	单次运行，冻结测试集划分	pass
M-SYNTH，冻结	仅元数据	0.5000	单次运行，冻结	通过（0.50 by

表5：**已标注的 KAU-BCMD 文件夹（本地副本）中按类别统计的图像尺寸。**

类别	图像	不同的图像尺寸	尺寸（数量）
BI-RADS 1（阴性）	1,865	3	2816×3584(1,552) · 1284×577(237) · 824×578 (76)
BI-RADS 4（阳性）	102	2	467×578 (98) · 482×578 (4)
BI-RADS 5级（阳性）	24	24	24 张图像中的每一张尺寸均不相同
类别之间共享	~	0	没有任何阳性图像具有与任何图像相同的尺寸阴性图像

在两个队列和合并数据中均作为对照组（第 3.4 节），并仍被计为几何方面的失败。

3.2 KAU-BCMD：类别是文件头的函数

在本地副本中，所有超过 5 百万像素的图像均为 BI-RADS 1，共 1,552 张；而所有 BI-RADS 4 图像均为 0.27 百万像素。阴性图像有三种尺寸；阳性图像尺寸较小，其中 BI-RADS 5 图像各自具有独立尺寸。这种模式正是对两个类别采用不同导出或裁剪流程可能产生的结果；我们未观察文件的制作过程。类别可以从文件头中判定，几何控制得到的曲线下面积为 1.0000。

图像内容无法解释这种区分。在显示尺寸匹配的情况下进行目视检查可见，两个类别均为完整乳房的头尾位和内外侧斜位视图，而不是病灶裁剪图（图1）。不同之处在于文件的生成方式：在初步分析中，阳性图像每像素携带的高频能量系统性地高于阴性图像。

公开Kaggle档案呈现出相同的关联。我们将该档案与解压后的副本逐个文件进行了比较：全部 2,378 张图像的文件夹、文件名、尺寸和字节大小均一致。其 BI-RADS 1 图像的尺寸和数量与表5完全相同。其 BI-RADS 4 图像为 1267×578(98) 和 1282×578(4)：数量和高度与本地副本相同，但宽度恰好大 800 像素。因此，本地文件夹是同一批图像的另一种导出结果，并非对该档案的转录；本地副本是由 DICOM 导出生成的 PNG文件夹。24 张 BI-RADS 5级图像的尺寸仍各不相同，宽度均为 183–477 像素。档案中的不同类别之间也没有共享任何尺寸。

档案中的数值低于本地的 1.0000，原因明确。一名患者包含全部四张宽度为 1282 像素的 BI-RADS 4 图像。1282 的宽度介于阳性宽度 1267 和阴性宽度 1284 之间，因此，在任何将该患者划入测试折的拆分中，分类器都从未见过这一宽度，并会将这四张图像排在阴性图像一侧。它们在每种拆分中都会被错误排序，其他图像则不会；使用精确图像尺寸进行查找且不进行拟合，得出的结果为 1.0000。在本地导出中，BI-RADS 4 图像的宽度为 467 像素，并且

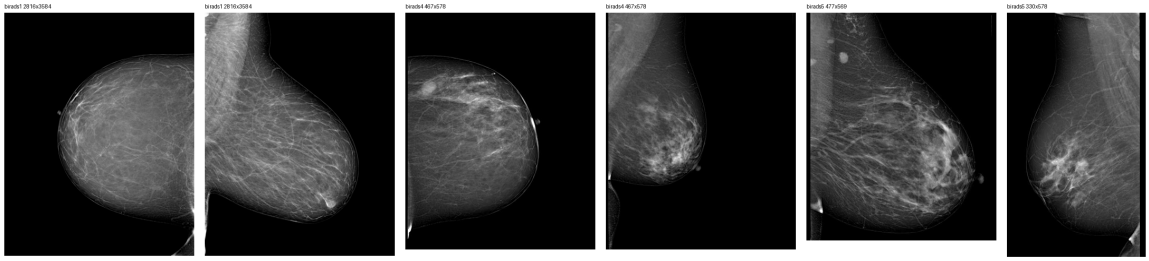


图 2：**本地 KAU-BCMD 副本中按类别展示的样本图像**。已标注文件夹中三个 BI-RADS 类别各选取两张图像，以相同的显示尺寸呈现，每个面板上方标注原始像素尺寸（BI-RADS 1：2816×3584；BI-RADS 4：467×578；BI-RADS 5：477×569 和 330×578）。这两类图像均为完整乳房的头尾位和内外侧斜位视图，而非病灶裁剪图。在此副本中，阴性类别包含三种图像尺寸（2816×3584，$\mathrm{n} = 1,552$；1284×577，$\mathrm{n} = 237$；824×578，$\mathrm{n} = 76$）；BI-RADS 4 图像包含两种（467×578，$\mathrm{n} = 98$；482×578，$\mathrm{n} = 4$）；24 张 BI-RADS 5 图像的尺寸各不相同，且没有任何阳性图像与阴性图像的尺寸相同；仅使用宽度、高度、百万像素和宽高比作为输入的分类器，在按患者分组的五折交叉验证中区分类别的曲线下面积为 1.0000。图像来自阿卜杜勒-阿齐兹国王大学乳腺癌乳腺X线影像数据集（KAU-BCMD；Alsolami 等人 [25]），采用知识共享署名 4.0 国际许可协议授权，并已调整大小以便展示。

表 6：**对公开的 KAU-BCMD 存档进行几何控制，比较 BI-RADS 1 与 4/5**。所有行均为 20 次重复患者分组五折划分的均值。

设置	图像	曲线下面积均值	SD	20 次数据划分的范围	结论
所有患者	1,991	0.9790	0.0082	0.9712–0.9995	fail
排除这 21 名患者同时拥有两类图像类别	1,903	0.9397	0.0320	0.8791–0.9980	fail
仅按病例编号分组（64 个病例编号在两个年份前缀下重复出现）	1,991	0.9801	0.0051	0.9713–0.9882	fail

表 7：**CMMD 按子集合划分的统计，数据来自标签文件。**

子集合	良性	恶性	总计	恶性比例	分子亚型 记录值（比例）
D1	544	563	1,107	0.508582	0.000000
D2	12	753	765	0.984314	0.979085

远低于所有阴性图像的宽度，因此单一阈值即可区分类别，这与在那里测得的 1.0000 一致。

预处理并未消除这种关联。在本地副本中，将所有内容调整至 224×224 会消除几何通道，但仍留下可区分类别的纹理差异，得分为 0.8625；在调整尺寸前先将短轴均衡至统一的 467 像素，得分为 0.9297；而分辨率匹配子集的得分为 1.0000，因为实际上不存在真正匹配的图像：所谓“匹配”的阴性图像为 0.48 兆像素，而阳性图像为 0.27 兆像素。在公共档案中，三种补救方法的结果分别为：调整尺寸后 0.8465（标准差 0.0110，范围 0.8233–0.8641）、短轴均衡化后 0.8529（标准差 0.0097，范围 0.8342–0.8743），以及 0.5 兆像素以下子集的 0.9800（标准差 0.0112，范围 0.9649–1.0000）；该子集包含来自 28 位患者的 76 张 BI-RADS 1 和 24 张 BI-RADS 5级图像。在本地副本中，短轴均衡化的得分高于直接调整尺寸。这一观察结果未能在公共档案中重现；两种方法的得分相差 0.006，处于一个标准差以内，因此本文不对此作出主张。我们的解释是——但我们并未进一步验证——重采样无法让两个类别拥有相同的处理历史：它无法恢复阴性图像在统一尺度下缺失的信息，也无法为其添加阳性图像所带有的压缩特征。

如果文件头不包含信息，规模将会如何。 本地副本包含501名患者和1,000个乳房，其中60个属于乳腺影像报告和数据系统4–5类（分别为50个和10个），另有938个属于BI-RADS 1类。在25%的测试集划分下，阳性测试乳房约有15个，因此可检测到的最小曲线下面积差异约为0.15，远高于计划基准测试所针对的0.02。即使文件头不包含信息，本地副本也无法支持该基准测试。938 / 50 / 10这一组计数不包括2个文件夹标签相冲突的病例-侧别单元；若强行给这2个单元添加标签，则计数为939 / 51 / 10。两组计数均予以保留。

缺失的类别。 本地副本中没有BI-RADS 0和乳腺影像报告和数据系统2类，其乳腺影像报告和数据系统3类文件夹中有一个文件不符合命名约定。公共档案中同样没有BI-RADS 0或乳腺影像报告和数据系统2类文件夹；其乳腺影像报告和数据系统3类文件夹中有387幅图像。

3.3 CMMD：子集合前缀可预测诊断

D1 中 50.9% 为恶性，D2 中 98.4% 为恶性；D2 中 97.9% 的记录标注了分子亚型，而 D1 中没有任何记录标注。仅凭前缀即可预测恶性肿瘤，曲线下面积为 0.7753（根据标签文件重新计算的完整精度值为 0.775303）。该数据集的作者描述了这种两部分结构 [26]。我们衡量的是它所造成的影响：任何在未控制子集合的情况下合并 D1 和 D2 进行的分析，都会继承一个曲线下面积为 0.7753 的信号通道。

几何控制在 CMMD 上的结果恰好为 0.5000，而清单揭示了原因：全部 5,202 张图像的像素尺寸均为 2294×1914，因此几何特征向量是常量，按构造，对照组的表现就等同于随机机会水平。位深并不统一：

表 8：**CMMD D1 上的标识符顺序控制**。唯一特征是患者标识符中的序列号。此对照是在事后添加的，属于探索性分析。

评估	乳房	AUC
冻结的 D1 验证集划分	171	0.9700
冻结的 D1 测试集划分	331	0.9926
20 次重复的患者分组五折划分	1,107	均值 0.9871（标准差 0.0009；范围 0.9852–0.9892）

其中 5,200 张图像以 8 位存储，另有 2 张以 16 位存储。此前对两个序列样本进行的一次检查曾记录档案中的图像均为 8 位；这两项观察结果均予以保留。

在合并数据集上，纹理控制以 0.6275 的结果未通过 0.60 阈值；正是这次失败触发了对子集合的调查。在限定于 D1 并重新冻结数据划分后，几何为 0.5000，纹理为 0.5810，二者均低于停止阈值。我们将 D1 的纹理值视为下限而非缺陷，这一判断基于一个未经检验的假设，即全局纹理统计量与乳腺密度协变；据此，在 D1 上评分达到或低于 0.58 的模型，除了四个全局统计量所提供的信息之外，并未展示出任何额外能力。

仅年龄一项在合并数据集上的评分为 0.6832，在 D1 上为 0.6744。这两个数值均作为上下文予以报告，且都不会使研究停止：CMMD 是一个诊断队列，其中年龄与恶性肿瘤相关，这是人群的一项属性。不过，对于阅读任何 CMMD 结果的读者而言，该数值仍然有用，因为模型必须超过这一性能，才能证明其使用了图像。

3.4 CMMD D1：标识符顺序可预测诊断（事后控制）

按序列号十分位分组后，从最低序列号到最高序列号，各组的恶性比例依次为 0.54、0.36、0.89、0.43、0.00、0.00、0.00、0.88、1.00 和 0.98。因此，发布的标签文件中，同一诊断的记录长段连续排列；仅凭记录的位置，分类器几乎就能完美还原标签。

由此得出三点。第一，仅限于 D1 的 CMMD 通过了已注册的几何和纹理控制（0.5000 和 0.5810），但未通过这项控制；它并非干净的阴性对照。第二，这一结果在本文中表明，通过已注册的控制并不等于获得证书。第三，发布的门控现已规定第五项控制，即标识符顺序，并说明该项是在事后添加的。

标识符并不是图像的属性，仅凭像素输入的模型无法直接读取标识符。此前未检验任何图像属性是否会与序列顺序协变，例如通过采集时期产生协变。作者在另一项基准测试中发现了同样的泄漏，导致其中使用 CMMD D1 的一个分支被撤回。

CBIS-DDSM 上的同一对照组（同样为事后分析）仍低于停止阈值：钙化队列的均值曲线下面积为 0.5308（1,511 张图像，753 名患者），肿块队列为 0.4777（1,592 张图像，892 名患者），合并后为 0.5399（3,103 张图像，1,566 名患者）；60 个数据划分值全部不超过 0.5570。在官方划分上，其结果为 0.5299（钙化，284 张测试图像）和 0.4451（肿块，361 张测试图像）。按序列十分位统计的恶性比例在钙化队列中并不均匀（第 1–9 个十分位为 0.34 至 0.46，第 10 个十分位为 0.66），但该

表9：**对 CBIS-DDSM 进行几何控制**。20 次重复划分的均值是判定统计量，该指标是在查看结果后选定的；旁边列出了官方单次划分的数值。

队列	图像 / 患者	平均曲线下面积 20 个数据划分（范围）	结论 mean	仅限官方划分
钙化	1,261 / 635	0.6876 （0.6669–0.7040）	fail	0.5625（37 张测试图像）；低于 stop
Mass	1,592 / 892	0.6349 （0.6243–0.6438）	fail	0.5884；低于 stop
合并	2,853 / 1,460	0.6543 （0.6419–0.6643）	fail	0.6182（31 重叠患者已移除）；高于停止阈值

标识符本身无法区分类别。因此，CBIS-DDSM 标签文件不存在 CMMD 的排序缺陷，而 CBIS-DDSM 的结论仅基于几何。

3.5 CBIS-DDSM：高于停止阈值的几何—标签关联

每张全视野图像均根据其自身的 DICOM头进行定位（读取了 10,229 个 DICOM对象；根据序列描述和患者标识符识别出 2,853 张全视野乳腺X线影像），并与官方病例描述文件关联。标签为图像级病理学：如果图像上任一标注病灶为恶性，则标记为恶性；否则标记为良性。保留了官方的按队列划分，并将两个队列中的患者分组处理。

这种关联弱于 KAU-BCMD 中的关联，也不是查找表所致：在合并队列中，两种类别共有 122 种图像尺寸，2,853 张图像中出现了 2,448 种不同尺寸。在全部 60 次重复划分中，关联均高于停止阈值。不过，结论仍取决于看到结果后作出的选择。在单次官方划分中，肿块队列的得分为 0.5884，钙化队列为 0.5625；后者仅基于 37 张测试图像，且两者均低于停止阈值。移除 31 名重叠患者后，合并队列的得分仍以 0.6182 超过停止阈值。因此，我们将这一结果描述为：在重复划分统计量下，几何控制未能通过，而非数据集存在缺陷。本研究未调查其作用机制，也未将这种关联归因于任何特定的采集、数字化或整理步骤。

副本的一个特征限定了钙化结果的范围。钙化测试集划分中几乎没有这类图像，在 284 个官方乳房视图中仅占 37 个，因此钙化这一行主要描述官方训练图像。

这两个官方队列还会与官方数据划分相互作用。将钙化队列和肿块队列合并后，有 31 名患者同时出现在官方训练集和官方测试集划分中，其中 18 名同时属于钙化训练集和肿块测试集，另有 13 名情况相反；此外，还有 29 个乳房视图同时存在于两个数据划分中；而在任一单独的队列内部则不存在重叠。任何在保留官方数据划分的同时合并两个队列的分析，都会将这些患者置于两侧。其他研究者已量化了由评估方案导致的 CBIS-DDSM 患者级数据泄漏 [15]。

3.6 通过对其运行控制变量的数据集

M-SYNTH 通过了现有的两项对照检验：全图物理特征得分为 0.5070，而仅使用元数据、乳腺密度和剂量标签时得分为 0.5000，后者按构造即为 0.50。

INbreast 通过了几何控制，均值为 0.4863（范围为 0.4367–0.5153）。分析的 379 张图像仅有两种尺寸，2560×3328 和 3328×4084，且这两种尺寸均出现在两个类别中。

现有的 VinDr-Mammo 子集以 0.5732 通过（范围为 0.5654–0.5797），这是最接近停止阈值的一次通过；两个类别共有 21 种图像尺寸。这个数值反映的是 20,000 张已发布图像中的 11,965 张；按 BI-RADS 分类划分的每个单元中，现有图像占比均在 57.9% 至 69.7% 之间。它并不代表完整的发布版本，而该子集可能经过了控制变量无法识别的筛选。

对 INbreast 和 VinDr-Mammo 仅运行了几何控制，而 M-SYNTH 仅有两项控制变量。通过一项或两项控制并不能证明数据集毫无问题，正如第 3.4 节所示，有一个数据集通过了两项控制。

3.7 已注册但未完成的项目

KAU-BCMD 的第二个发布方路径未经过核查，未通过机构账户获取档案，并且尚未确定哪个导出版本生成了本地文件夹，也未确定 800 像素的宽度差异是由裁剪、填充还是不同发布版本造成的。

RSNA 挑战赛数据集 [38] 和 BreastMNIST [39] 未进行门控。在 CBIS-DDSM、INbreast 或 VinDr-Mammo 上，纹理、批次/站点、协变量和位深控制均未运行。

注册声明称，如果后续五个数据集全部通过，泛化主张将被撤回。在已注册的单次划分设计下，CBIS-DDSM 的肿块队列和钙化队列均低于停止阈值，只有合并队列（这是我们构建的一种配置）超过了该阈值。因此，该条件是否满足取决于事后统计量，或取决于是否将合并配置计入其中。我们仅主张，在第三个数据集的某些配置上，这些控制变量返回了高于停止阈值的数值。

初始控制变量未在重复划分上重新运行。两个纹理值 0.6275 和 0.5810 分别位于 0.60 停止阈值的两侧，其离散度可能改变结论；0.7753 的批次前缀则不存在这一弱点。

未测量未通过的控制所带来的下游代价，也就是传统分类器在存在和不存在泄漏通道时曲线下面积的变化。

---

4 讨论

4.1 控制变量的代价及其本质

几何控制对每个文件读取四个整数，不解码任何像素。对于约有 2,000 张图像的数据集，例如包含 1,991 张图像的 KAU-BCMD 清单，分类器拟合只需几秒。若文件存储在云存储中，读取文件头的耗时更长：对于一个包含 10,000 个对象的数据集，需要几分钟（笔记本在处理 10,229 个

CBIS-DDSM 对象时约运行了四分钟，而处理 VinDr-Mammo 子集时约需 25 分钟）。纹理控制需要解码像素，因此速度较慢。在出现失效的三个数据集上，控制项在本地 KAU-BCMD 副本上返回 1.0000，在公共档案上返回 0.979，在合并的 CBIS-DDSM 上按重复划分统计量计算为 0.654，并通过批次控制在 CMMD 上返回 0.7753。与基于受污染队列构建基准测试的代价相比，控制项的代价很小。这正是应该在数据接收时运行这些控制项，而不只是出现异常迹象时才运行的理由。

该设计的两项特性与其实现方式彼此独立。对照组从构造上就存在不足：它们接收的输入本不应包含诊断信息，因此高分很难被解读为泄漏以外的任何东西。这使它们比性能上限或学习曲线更有信息量，因为后两者都可以用模型容量来解释。协变量控制则只作报告而不强制执行，因为如果某个门控因年龄在诊断队列中能够预测恶性肿瘤而停止一项研究，那么它衡量的将是流行病学。

4.2 调整尺寸并非防御手段，且作用机制各不相同

在 KAU-BCMD 上，将图像调整为标准输入尺寸并未消除这种关联。它去除了几何通道，却留下了纹理差异，足以区分类别：在本地副本上为 0.8625，在公共档案上为 0.8465；先统一短轴长度也无济于事，在两个副本上均是如此（本地为 0.9297，公共档案上为 0.8529）。在该数据集中，混杂因素由图像的信息内容承载，而不仅仅是图像尺寸，因此，匹配尺寸只能解决读者可见的问题，却没有消除其成因。

CMMD 展示了相同后果下的另一种机制。该档案在维度上是统一的，在几何控制上得到机会水平的结果。问题在于，恶性肿瘤比例分别为 50.9% 和 98.4% 的两个子集合被共同分布，因此仅凭标识符前缀就能达到 0.7753 的曲线下面积。几何控制和批次控制都无法发现对方的数据集，这正是需要不止一种控制的理由。CBIS-DDSM 展示了第三种模式：没有任何单一维度能够区分类别，但在重复划分下，尺寸与形状的联合分布与标签的关联超过了停止阈值。若检查只寻找某一类别独有的图像尺寸，而不对分类器进行评分，就不会发现这一问题。

4.3 通过并不等于证书

CMMD D1 是一个具有启发性的案例。合并数据集失败后，它被采用为修正后的队列；它通过了两个已注册的停止阈值控制，并在加入标识符顺序控制之前一直作为工作队列。一个此前未被预料到的控制因素——标识符中的序列号——随后返回了 0.9926 的测试曲线下面积。控制措施只能检测其作者预期到的通道；单个未经调优的学习器只能给出泄漏的下限，而不是上限。对表4中每个“通过”的恰当解读是：运行的控制措施未检测到泄漏；而对于 INbreast 和 VinDr-Mammo，这意味着只运行了一项控制。

标识符顺序控制是在事后添加的，而在结果已知后再添加控制本身就是一条分叉路径。因此，我们将其报告为探索性分析；我们并不声称，既然四个控制变量不够，五个控制变量就足够。

4.4 对这些数据集用户的启示

表3所标示的公开Kaggle档案中存在KAU-BCMD关联，因此，任何使用该档案训练的研究都与此相关，读者也可以在自己的副本上进行核查。我们不会将这一说法推广到该数据集的其他分布。CMMD的发现是已发布标签文件本身的属性：每一行标识符中都包含子集合前缀和序列号；任何汇集D1和D2的分析，或以任何方式使用D1内部记录顺序的分析，都会继承相应的信息通道。对于CBIS-DDSM，在重复划分下，几何与病理学的关联高于停止阈值，但在官方肿块与钙化划分中则不存在这种关联；按官方划分汇集两个官方队列，会导致31名患者同时出现在划分的两侧。

实际建议应当具体且可检验：任何使用公共乳腺X线摄影数据集进行训练的论文，都应报告对照组AUC；在进行汇集时，说明子集合的构成；合并来自不同队列的官方划分时，检查患者重叠情况；并将记录顺序和标识符视为潜在的标签载体。

4.5 量子与经典研究计划中的上下文

本文之所以应与一组量子与经典比较研究并列，仅仅是因为对照组的设计旨在保护这些比较。本文与所提出的量子优势框架 [36] 的关系仅限于一个观察：该框架要求验证输出的正确性，而如果数据集的标签可以从其文件头中读取，那么这种验证对每个实验分支都将失效，并且这一缺陷先于计算过程本身。该框架由与量子计算供应商有关联的作者提出，我们不采纳其任何预测。本文没有报告量子优势、加速或量子增强性能，没有将任何量子结果作为任何优势的证据，推理路径中也不存在量子组件。

---

5 局限性

1. **KAU-BCMD 档案是两条注册途径之一，其来源信息并不完整**。 我们根据文件名、大小和文件成员日期，在一份托管于 Kaggle 的存档中确认了这种关联。下载时未记录托管页面、上传者和版本。该档案的文件名与数据集论文数据可用性声明中所列 Kaggle 页面的标识符相符[25] (https://www.kaggle.com/asmaasaad/king-abdulaziz-university-mammogram-dataset)，这与数据集作者自行存档的说法一致，但不足以证实这一点；不能排除由第三方重新上传的可能性。第二条注册途径未经核查。

2. **对 KAU-BCMD 本地副本的分析未经注册**。 这是一项临时诊断，门控是在事后正式确立的，因此结果属于探索性结论。

3. **注册为内部注册**。这是一份带日期的书面方案，附有偏差日志，与代码一同发布，而不是公共注册库中的记录。4. **批次/站点控制和标识符顺序控制是事后增加的**。CMMD子集合混杂因素是在纹理控制失败后通过调查发现的，标识符顺序控制则是在注册门控通过 CMMD D1 后增加的，随后应用于 CBIS-DDSM。此处发布的五项控制门控在设计上因此部分具有回溯性。

5. **初始结果是单次运行、使用一个随机种子得到的**。这些结果没有离散度可供使用，且 0.6275 和 0.5810 这两个纹理值分别位于停止阈值两侧。6. **后续结果的判定统计量是在看到结果后选择的**。后续结果带有 20次划分离散度，其结论取决于均值。在单次官方划分上，CBIS-DDSM 肿块队列得分为 0.5884，钙化队列得分为 0.5625（37 张测试图像），均低于停止阈值；而在移除 31 名重叠患者后，合并队列超过停止阈值，得分为 0.6182。7. **对于 CBIS-DDSM，几何控制的前提较弱**。在一个裁剪至乳房区域的数字化胶片数据集中，图像尺寸可能部分反映乳房大小、数字化仪或站点的影响，而这些因素均未经过测试；因此，对于 CBIS-DDSM，几何不携带诊断信息这一前提，比固定格式数字数据集中的该前提更弱。8. **单个未经调优的分类器只能给出泄漏的下界**。控制失败就意味着失败；控制通过仅表示该学习器在这些特征上没有发现任何问题。门控是单侧筛查，通过并不等于获得证书，CMMD D1 的结果便说明了这一点。9. **不同数据集之间的方案并不统一**。本地 KAU-BCMD 副本仅进行一次交叉验证，CMMD 和 M-SYNTH 在冻结划分上各评分一次，而后续分析则在 20 个重复交叉验证划分上进行评分。可用的控制因记录的元数据不同而有所差异，且只有几何控制在 CBIS-DDSM、INbreast 和 VinDr-Mammo 上运行。10. **生成初始 KAU-BCMD 和 CMMD D1 数值的脚本不可用**。发布的门控脚本是根据其书面规范重构的，并非生成这些数值的代码。原始输入和测量输出均已发布；对于相关记录，二者之间的转换以文字形式记录。后续结果由已发布的笔记本端到端生成，并与其输出一同保存。11. **合并 CMMD 数据划分的注册校验和无法核验**。在方案修订为 D1 时，合并的数据划分文件被替换，无法恢复。发布版本同时报告了注册校验和与实际校验和。12. **六个数据集不足以作为一般性建议的基础**。KAU-BCMD 和 CMMD 是发现首次失败的两个数据集，因此样本的这部分并不独立于该发现。CBIS-DDSM、INbreast 和 VinDr-Mammo 则是在检查之前就已在注册中列出的。13. **KAU-BCMD 有三种留档规模，且彼此不一致**。数据集论文报告了 1,416 个病例和 5,662 张图像 [25]；IEEE DataPort 中的数据集记录列出了 1,521 个病例和 6,109 张图像；而包含该数据集的基准数据汇编 [33] 列出了 1,416 个病例和 2,206 张 JPG 图像。我们通过数据集论文的摘要确认了第一组数据，并通过该汇编文本的摘录确认了第三组数据。第二组数据是我们在初次数据集核验期间转录的，尚未

已针对本文重新核查。分歧仍未解决。这不影响任何对照组曲线下面积，但关于整个数据集的任何表述，其分母都尚不确定。

14. **CMMD 的位深并不统一**。对清单进行全面检查后发现，有 5,200 张图像为 8 位，另有 2 张为 16 位；而此前抽样检查记录的结果是所有图像均为 8 位。两种结果均予以保留，并在门控中加入位深检查。

15. **BI-RADS 是放射科医师评定的类别，并非病理学结果**。KAU- BCMD、INbreast 和 VinDr-Mammo 的标签与 CMMD 和 CBIS-DDSM 的病理学标签属于不同类型；BI-RADS 0 则表示工作流程中的一种操作，而非组织状态。

16. **后续副本并不完整，或仅有部分特征得到描述**。VinDr-Mammo 占发布版本的 59.8%；CBIS-DDSM 副本缺少 284 个官方钙化测试乳腺视图中的 247 个；INbreast 的访问协议未归档。每项结果均仅描述所分析的副本。

17. **KAU-BCMD 中的患者身份是根据文件名推断的**。64 个病例编号在两个年份前缀下重复出现；仅按病例编号分组会使档案的曲线下面积改变 0.001（表 6），但发布方并未记录患者结构。

18. **后续复跑采用了固定的纹理定义**。初始分析并未以数值形式定义这些纹理，因此无法证实本地与公开纹理的曲线下面积是以相同方式计算的。

19. **未评估**。这些控制变量能否检测几何、纹理、批次、标识符顺序和单一协变量之外的混杂因素；乳腺摄影以外的任何数据集；以及未通过的控制所造成的下游性能代价。

20. **阈值选择**。只有在接近阈值时，结论才会受到阈值影响。将 24 个停止角色结果分别按 0.55、0.65、0.70 和 0.75 重新评分后，未通过的结果数量会从阈值为 0.60 时的 14 个变为 16、12、10 和 10 个。在该范围内，两个 KAU-BCMD 副本和合并后的 CMMD 均未通过，与阈值无关，因为它们各自至少有一个数值达到 0.7753。三个数据集级判定确实会受阈值影响：CBIS-DDSM 在阈值不高于 0.6876 时因几何未通过，在 0.70 时则通过；VinDr-Mammo 在 0.60 时以 0.5732 的均值通过，但在 0.55 时会未通过；CMMD D1 在 0.60 时通过其注册控制，但在 0.55 时会因其纹理值 0.5810 而未通过。所有数值均已完整报告，因此偏好其他阈值的读者可以直接重新评估各数据集。

---

6 结论

训练前运行的文件级控制，在 KAU- BCMD 的副本和合并后的 CMMD 上得出了远高于机会水平的数值；在 CBIS-DDSM 上，按照事后选定的统计量计算，结果也高于 0.60 的停止阈值。一个通过了注册控制的队列，后来却未能通过一项新增的控制。这些控制成本低廉，只能检测其设计目标所针对的问题；无论结果是否显著，都值得与基于这些数据集训练的任何模型一并报告。

《数据》可用性

支撑本文的派生表已存入 Zenodo 公共代码仓库，doi:10.5281/zenodo.24372683。该存档不包含乳腺摄影图像。图1复现了 KAU-BCMD 中的六幅图像，依据该数据集的知识共享署名 4.0 国际许可协议使用。

该存档包含：gate_summary.csv（每个对照结果及其来源）；thre shold_sensitivity.csv 和 threshold_sensitivity_summary.json（在0.55至0.75的阈值范围内对每个结论重新评分，以及近似随机水平的划分差异）；gate_results_d1.csv 和 gates_msynth.csv（CMMD D1和M-SYNTH对照结果）；gate_res ults_d1_identifier_order.csv 和 gate_results_d1_identifier_order_splits.csv（对CMMD D1进行的事后标识符顺序对照、固定划分值和每个划分的20个值）；gate_results_cbis_identifier_order.csv 和 gate_results_cbis_identifi er_order_splits.csv（对CBIS-DDSM进行的相同对照，按队列和合并结果列出，并包含每个划分的60个值）；cmmd_labels.csv、cmmd_manifest.csv 和 cmmd_batch_table.csv（CMMD标签表、逐图清单和子集表）；split_train.csv、spli t_val.csv 和 split_test.csv（固定的CMMD D1数据划分）；kau_bcmd_labelled_inventor y.csv、kau_bcmd_local_inventory_redacted.csv 和 kau_geometry_table.csv（本地KAU-BCMD清单和尺寸表；本地清单中的导出文件夹名称已替换为不透明标签）；kau_bucket_inventory.csv（公开的 KAU-BCMD 存档中全部2,378幅图像的逐图宽度、高度、位深、压缩、文件大小和SHA-256）；cbis_split_overlap_patients.csv（同时出现在两个官方划分中的31名CBIS-DDSM患者）；checksums.sha256和verification_log.md（校验和及重新推导记录）；以及 01_preregistration.md（包含带日期偏差日志的内部方案）。逐图清单仅针对公开的 KAU-BCMD 副本提供；CBIS-DDSM、INbreast或VinDr-Mammo均未制作此类清单。

源数据集由读者向其提供方获取。KAU-BCMD[25] 以托管于 Kaggle 的存档 king-abdulaziz-university-mammo gram-dataset.zip 形式获取，记录的许可协议为知识共享署名 4.0 国际许可协议。文件名与数据集论文 [25]中提供的 Kaggle 页面相匹配（ https://www.kaggle.com/asmaasaad/king-abdulaziz- university-mammogram-dataset）；版本和下载日期未予记录。CMMD [26, 27] 由癌症影像档案库 [28] 依据知识共享署名 4.0 国际许可协议分发（https://doi.org/10.7937/tcia.eqde-4b16）。CBIS-DDSM [29] 由癌症影像档案库依据知识共享署名 3.0 及该档案库的数据使用政策分发。VinDr-Mammo[31] 由 PhysioNet 依据 PhysioNet受限健康数据许可证 1.5.0 分发，并采用凭证访问。INbreast [30] 根据访问协议应请求从其提供方获取；截至本文撰写时，该协议未由作者存档。M-SYNTH [32] 由美国食品药品监督管理局作为监管科学工具 RST24AI06.01 依据 CC0 分发。

代码可用性

代码已存放在同一代码仓库中（doi:10.5281/zenodo.24372683）。其中包括 acceptance_gates.py（可复用形式的门控）；patchfeat.py（原始 M-SYNTH 控制代码）；verify_numbers.py（根据表格重新推导文中引用的计数）；threshold_sensitivity.py（在其他阈值下重新评定每项结论）；cmmd_d1_id

entifier_order.py 和 cbis_ddsm_identifier_order.py（针对 CMMD D1 和 CBIS-DDSM 的事后标识符顺序对照；后者会读取四个公开的 CBIS-DDSM 病例描述文件，读者可从档案中获取这些文件）；make_fig2_gate_pan el.py（图 2）；五个附有输出的笔记本：kau_bcmd_comparison.ipynb、kau_bcmd_preprocessing_controls.ipynb、cbis_ddsm_geometry.ipynb、inbreast_geo metry.ipynb 和 vindr_mammo_geometry.ipynb；run_all.sh；以及 requirements.txt、requirements_lock.txt 和 ENVIRONMENT_MANIFEST.md（软件版本和机器配置）。

acceptance_gates.py 按书面规定实现了该程序，并非用于生成最初 KAU-BCMD 和 CMMD D1 数值的代码；生成这些数值的脚本目前不可用。标识符顺序控制（cmmd_d1_ identifier_o rder.py）在已发布的划分文件上使用相同的分类器，并在相同划分上重现记录的年龄值，以检查方案。

作者贡献

A.T.：研究、软件、数据整理、验证、撰写——审阅与编辑。R.S.：研究、验证、撰写——审阅与编辑。M.Y.E.H.：概念化、方法论、形式分析、监督、撰写——初稿、撰写——审阅与编辑。

伦理声明

本研究仅使用了公开可用的、去标识化或模拟的数据集（KAU-BCMD、CMMD、CBIS-DDSM、INbreast、VinDr-Mammo 和 M-SYNTH）。作者未开展人体受试者研究，未收集患者数据，也无需获得伦理批准。

人工智能使用声明

在作者的监督下，基于大语言模型的助手用于编辑和重组作者撰写的文本、核查参考文献和引用数字，以及准备投稿文件。所有内容均经作者审阅并批准，作者对稿件承担全部责任。

更广泛影响声明

此处所述的控制措施旨在阻止模型在其标签可从非诊断信息中恢复的数据集上进行训练和报告，从而保护患者、临床医生和监管机构，避免其面对与疾病毫无关系的性能声明。可能存在两种误用。控制措施未通过，可能被解读为表明某个数据集或其创建者存在过错；但这些发现描述的是特定发布版本及标签文件的属性，其中有若干是事后选择或添加的；而控制措施通过，则可能被解读为质量证书，但本文明确否定了这种解读。本文不发布任何图像或新的患者数据；存档内容仅包含代码和派生表。

致谢与经费披露

作者感谢创建和维护本文所分析公共数据集的人员。本研究由 Qanary Inc. 内部资金支持。**利益冲突**。所有作者均为 Qanary Inc. 的员工，该公司正在开发乳腺摄影筛查软件。M.Y.E.H. 和 A.T. 是一项正在审查中的美国专利申请（编号：19/876,140，提交日期：2026年11月17日）中列名的发明人，该申请涉及输入完整性门控。作者声明不存在其他利益冲突。

参考文献

[1] Geirhos, R. 等. 深度神经网络中的捷径学习。《自然·机器智能》**2**, 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z[2] Lapuschkin, S. 等. 揭开聪明汉斯预测器的面纱，评估机器真正学到了什么。《自然·通讯》**10**, 1096 (2019). https://doi.org/10.1038/ s41467-019-08987-4[3] Torralba, A. & Efros, A. A. 对数据集偏差的无偏审视。载于*CVPR 2011*, 1521–1528 (2011). https://doi.org/10.1109/CVPR.2011.5995347[4] Zech, J. R. 等. 深度学习模型在胸部X线片中检测肺炎的可变泛化性能：一项横断面研究。《*PLOS*医学》**15**, e1002683 (2018). https://doi.org/10.1371/journal.pmed.1002683[5] Badgeley, M. A. 等. 深度学习利用混杂的患者与医疗保健变量预测髋部骨折。*npj Digital Medicine* **2**, 31 (2019). https://doi.org/10. 1038/s41746-019-0105-1[6] Winkler, J. K. 等. 皮肤镜图像中的手术皮肤标记与深度学习卷积神经网络黑色素瘤识别诊断性能之间的关联。《*JAMA*皮肤病学》**155**, 1135 (2019). https://doi.org/ 10.1001/jamadermatol.2019.1735[7] Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Re, C. 隐藏分层导致医学影像机器学习出现具有临床意义的失败。载于*ACM*健康、推理与学习会议论文集, 151–159 (2020). https: //doi.org/10.1145/3368555.3384468[8] DeGrave, A. J., Janizek, J. D. & Lee, S.-I. 用于放射影像 COVID-19 检测的人工智能选择捷径而非信号。《自然·机器智能》**3**, 610–619 (2021). https: //doi.org/10.1038/s42256-021-00338-7[9] Maguolo, G. & Nanni, L. 对 X 射线图像 COVID-19 自动检测方法的批判性评估。《信息融合》**76**, 1–7 (2021). https://doi.org/ 10.1016/j.inffus.2021.04.008

[10] Varoquaux, G. & Cheplygina, V. 医学影像机器学习：方法学失误及未来建议。*npj Digital Medicine* **5**, 48 (2022). https: //doi.org/10.1038/s41746-022-00592-y[11]Kapoor, S. & Narayanan, A. 机器学习科学中的泄漏与可复现性危机。*Patterns* **4**, 100804 (2023). https://doi.org/10.1016/j.pa tter.2023.100804[12] Leek, J. T. 等. 应对高通量数据中普遍且影响重大的批次效应。*Nature Reviews Genetics* **11**, 733–739 (2010). https://doi.org/10. 1038/nrg2825[13] Wachinger, C., Rieckmann, A. & P¨olsterl, S. 检测并纠正多中心神经影像数据集中的偏倚。*Medical Image Analysis* **67**, 101879 (2021). https://doi.or g/10.1016/j.media.2020.101879[14] Samala, R. K., Chan, H.-P., Hadjiiski, L. & Koneru, S. 机器学习中的数据泄漏风险：一项使用深度神经网络进行乳腺癌分类的研究。载于*Medical Imaging 2020: Computer-Aided Diagnosis* (eds Hahn, H. K. & Mazurowski, M. A.) 39 (SPIE, 2020). https://doi.org/10.1117/12.2549313[15] Elmejgari, C., Nadir, Y. & Qbadou, M. 乳腺X线摄影计算机辅助诊断中的可复现性鸿沟：量化患者级数据泄漏的影响。*Engineering, Tech-nology & Applied Science Research* **16**, 39844–39850 (2026). https://doi.org/10. 48084/etasr.20773[16] Liao, L. & Aagaard, E. M. 用于提升基于深度学习的乳腺癌诊断透明度的开放代码库：利用CBIS-DDSM数据。*Scientific Reports***14**, 27318 (2024). https://doi.org/10.1038/s41598-024-78648-0[17] Mabrok, M. A. 四视图乳腺X线摄影分类中的侧别级与患者级评估：基于CMMD数据集的全面基准测试。*Future Science OA***12**, 2681575 (2026). https://doi.org/10.1080/20565623.2026.2681575[18] Zufiria, B. 等. 深度学习模型开发中乳腺摄影数据集潜在偏倚的分析。载于*Lecture Notes in Computer Science*, 59–67 (2022). https://doi.org/10.1007/978-3-031-17721-7_7[19] Garrucho, L. 等. 乳腺摄影肿块检测中基于深度学习的领域泛化：一项大规模多中心研究。*Artificial Intelligence in Medicine***132**, 102386 (2022). https://doi.org/10.1016/j.artmed.2022.102386[20] Gebru, T. 等. 数据集数据表。*Communications of the ACM* **64**, 86–92 (2021). https://doi.org/10.1145/3458723[21] Wilkinson, M. D. 等. 科学数据管理与 stewardship 的 FAIR指导原则。《科学数据》**3**, 160018 (2016). https://doi.org/10.1038/sdat a.2016.18

[22] Mongan，J.、Moy，L. 和 Kahn，C. E. 医学影像人工智能检查清单（CLAIM）：作者和审稿人指南。《放射学：人工智能》**2**，e200029（2020）。https://doi.org/10.1148/ryai.2020200029[23] Oakden-Rayner，L. 探索大规模公共医学图像数据集。《学术放射学》**27**，106–112（2020）。https://doi.org/10.1016/j.acra.2019.10.006[24] Logan，J.、Kennedy，P. J. 和 Catchpoole，D. 对乳腺摄影机器学习数据集、其对 FAIR 原则的遵循情况及未来展望的综述。《科学数据》**10**，595（2023）。https://doi.org/10.1038/s41597-023-02430-6[25] Alsolami，A. S. 等。阿卜杜勒-阿齐兹国王大学乳腺癌乳腺X线影像数据集（KAU-BCMD）。《数据》**6**，111（2021）。https://doi.org/10.3390/data6110111[26] Cai，H. 等。包含活检确诊类型的在线乳腺摄影数据库。《科学数据》**10**，123（2023）。https://doi.org/10.1038/s41597-023-02025-1[27] Cui，C. 等。中国乳腺摄影数据库（CMMD）：一个包含活检确诊类型、用于乳房机器诊断的在线乳腺摄影数据库。癌症影像档案库https://doi.org/10.7937/tcia.eqde-4b16（2022）。[28] Clark，K. 等。癌症影像档案库（TCIA）：维护和运营公共信息存储库。《数字影像学杂志》**26**，1045–1057（2013）。https://doi.org/10.1007/s10278-013-9622-7[29] Lee，R. S. 等。用于计算机辅助检测和诊断研究的精选乳腺摄影数据集。《科学数据》**4**，170177（2017）。https://doi.org/10.1038/ sdata.2017.177[30] Moreira，I. C. 等。INbreast：迈向全视野数字乳腺摄影数据库。《学术放射学》**19**，236–248（2012）。https://doi.org/10.1016/j.acra.2011. 09.014[31] Nguyen，H. T. 等。VinDr-Mammo：用于全视野数字乳腺摄影计算机辅助诊断的大规模基准数据集。《科学数据》**10**，277（2023）。ht tps://doi.org/10.1038/s41597-023-02100-7[32] Sizikova，E. 等。用于比较评估不同乳房特征、病灶显著性和剂量范围下乳腺摄影人工智能的基于知识的计算机模拟模型与数据集。载于《神经信息处理系统进展》第*36*卷，37401–37412（2023）。https://doi.org/10.52202/075280-1626[33]Bhole，G.、Suba，S. 和 Parekh，N. Mammo-Bench：大规模乳腺摄影图像基准数据集。预印本，https://doi.org/10.1101/2025.01. 31.25321510（2025）。[34] Hajishafiezahramini，P.、Hamilton，M.、Meruvia-Pastor，O. 和 Kendall，E. 筛查乳腺摄影人工智能中的数据集来源特征与捷径学习：跨数据集案例研究。预印本，https://arxiv.org/abs/2607.15416（2026）。

[35] Zafari，Y. 等。MammoClean：通过数据集协调实现乳腺摄影人工智能的可复现性与偏倚感知。预印本，https://arxiv.org/abs/2511. 02400（2025）。[36] Lanes，O. 等。量子优势框架。预印本，https://arxiv. org/abs/2506.20658（2025）。[37] Hanley，J. A. 和 McNeil，B. J. 受试者工作特征（ROC）曲线下面积的含义与用途。《放射学》**143**，29–36（1982）。https://doi.or g/10.1148/radiology.143.1.7063747[38] Trivedi，H. M.、Vazirabad，M.、Kitamura，F. C.、Chen，Y. 和 Frazer，H. 用于 RSNA 筛查乳腺X线摄影癌症检测挑战赛的开源数据集。《放射学：人工智能》**8**，e250375（2026）。https://doi.org/10.1148/ryai .250375[39] Yang，J. 等。MedMNIST v2——用于二维和三维生物医学图像分类的大规模轻量级基准。《科学数据》**10**，41（2023）。https://doi.org/10. 1038/s41597-022-01721-8