深度学习和视觉基础模型在跨数据集评估中的非典型与正常有丝分裂分类基准测试

Sweta **Banerjee** [1], Viktoria **Weiss** [2], Taryn A. **Donovan** [3], Rutger A. **Fick** [4], Thomas **Conrad** [5], Jonas **Ammeling** [6], Nils **Porsche** [1], Robert **Klopfleisch** [5], Christopher **Kaltenecker** [7], Katharina **Breininger** [8], Marc **Aubreville** [1], Christof A. **Bertram** [2]

1 Flensburg University of Applied Sciences, Germany
2 University of Veterinary Medicine, Vienna, Austria
3 The Schwarzman Animal Medical Center, New York, USA
4 Diffusely, Paris, France
5 Freie Universität Berlin, Germany
6 Technische Hochschule Ingolstadt, Germany
7 Medical University of Vienna, Austria
8 Julius-Maximilians-Universität Würzburg, Germany

Abstract

非典型有丝分裂标志着细胞分裂过程中的异常，可以作为肿瘤恶性程度的独立预后相关标志。然而，由于其出现频率低、与正常有丝分裂在形态上的差异有时细微、病理学家之间评分的一致性较低以及数据集中类别不平衡等原因，识别它们仍然具有挑战性。基于乳腺癌非典型有丝分裂数据集（AMi-Br），本研究提出了一个全面的基准测试，比较了用于自动化非典型有丝分裂图像（AMF）分类的深度学习方法，包括基线模型、通过线性探测的基础模型以及使用低秩自适应（LoRA）进行微调的基础模型。为了严格评估，我们进一步引入了两个新的保留数据集——AtNorM-Br，即 TCGA 乳腺癌队列中的有丝分裂数据集，和 AtNorM-MD，一个多领域的来自 MIDOG++训练集的有丝分裂数据集。我们在域内 AMi-Br 以及域外 AtNorm-Br 和 AtNorM-MD 数据集中分别发现了高达 0.8135、0.7696 和 0.7705 的平均平衡准确性，其中 Virchow 系列基础模型基于 LoRA 的适应性表现尤为出色。我们的研究表明，虽然非典型有丝分裂分类是一个具有挑战性的问题，但可以通过利用最近在迁移学习和模型微调技术方面的进展来有效解决。我们将在本论文中使用的代码和数据在此 github 仓库中提供：`https://github.com/DeepMicroscopy/AMi-Br_Benchmark`。

Keywords

非典型有丝分裂，深度学习，基础模型，分类，基准测试，组织病理学，低秩适应

Article informations

Corresponding author: sweta.banerjee@hs-flensburg.de

1. 介绍

有丝分裂是细胞复制其遗传物质（DNA）然后分成两个完全相同的子细胞的过程。这就是细胞生长、修复受损组织并替换旧或磨损的细胞的方式。在分裂过程中，细胞可能会被显微镜观察到为一个 mitotic figure (MF)。量化肿瘤增殖对于评估许多人类和动物癌症至关重要，包括人类乳腺癌 (Fitzgibbons and Connolly, 2023; Kiupel et al., 2011; Louis et al., 2016)。在这个评估中的一个重要指标是有丝分裂计数，它反映了处于有丝分裂的细胞数量，并提供了关于肿瘤生长速度以及其组织学等级的重要信息。升高的有丝分裂计数与远处转移的风险增加及患者生存率降低有关 (Medri et al., 2003)。识别 MFs 大多仍然是一个手动任务，使其高度主观且耗时。此外，该任务还受到评分者间差异性的影响，使得用于自动检测的数据库创建变得复杂 (Meyer et al., 2005, 2009; Wilm et al., 2021)。深度学习模型已在有丝分裂检测和分类等任务中得到了广泛使用 (Aubreville et al., 2024; Veta et al., 2019; Aubreville et al., 2023b)，它

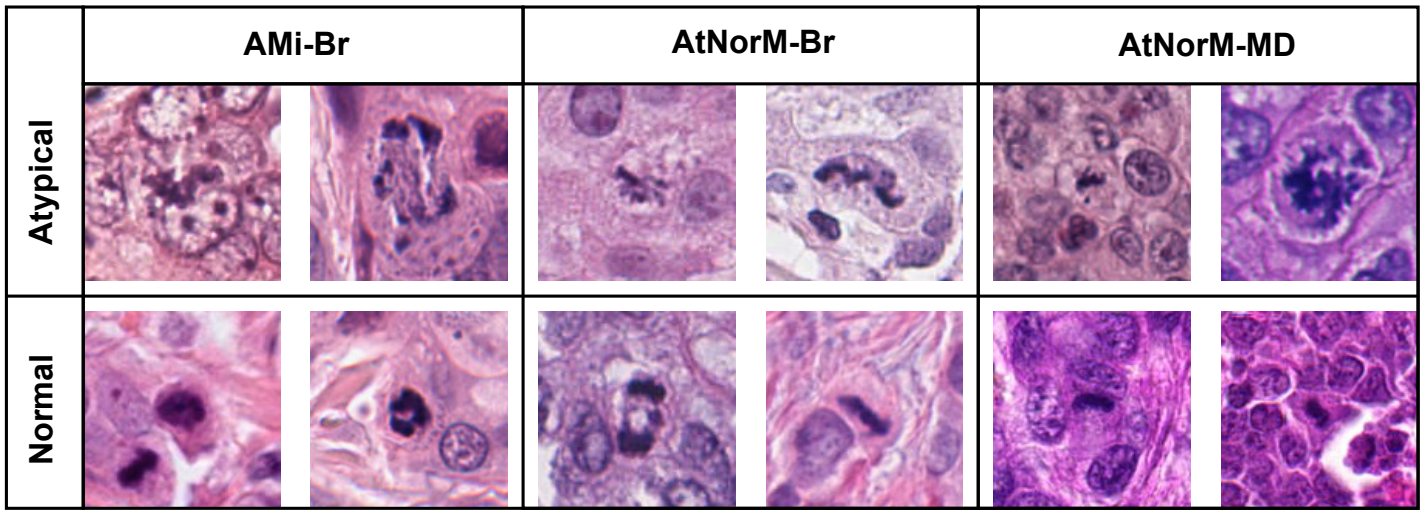

图 1: 显示本研究中包含的三个数据集中不典型和正常有丝分裂图的图像样本

们为这个问题提供了有希望的前景。

在肿瘤样本中，我们发现了两种类型的 MFs：正常和非典型。正常的 MFs 遵循典型的有序细胞分裂过程，包括前期、中期、后期和末期等阶段，在这些阶段染色体排列并分离以产生两个相同的子细胞 Donovan et al. (2021)。而非典型的 atypical mitotic figure (AMF) 则是经历异常细胞分裂过程的细胞，其特征是染色体分离错误或其他形态不规则现象，常在癌细胞中见到 (Donovan et al., 2021)。AMFs 的比率与乳腺癌 (Ohashi et al., 2018; Lashen et al., 2022) 和其他肿瘤类型 (Jin et al., 2007; Kalatova et al., 2015; Bertram et al., 2023; Matsuda et al., 2016) 的预后较差有关。最近，Jahanifar et al. (2025) 已在大规模评估中确认了比率和密度 AMFs 在各种肿瘤中的预后相关性。

然而，由于有丝分裂具有挑战性和不规则的形态学特征，对 AMF 进行正常或非典型分类时，低的一致性使得其识别变得复杂 (Bertram et al., 2023; Aubreville et al., 2023a)。深度学习模型有可能统一检测和分类有丝分裂，实现可重复性，并减少病理学家在这方面的负担。此外，基于自动化的深度学习分析允许大规模、数据驱动的预后相关标志物的识别，例如整个切片 (Jahanifar et al., 2025) 上的 AMFs 比率。

关于将深度学习应用于 AMFs (Aubreville et al., 2023a; Fick et al., 2024; Bertram et al., 2025) 的研究非常少。正常与非典型 MFs 分类的复杂性在于正常和非典型的有丝分裂图像 (Donovan et al., 2021) 形态重叠，以及 AMF 频率低导致的类别不平衡。尽管这是一个具有预后相关性的难题，但目前缺乏标准化的基准测试，使得难以对比各种方法的有效性。因此，开发合适的数据集并建立能够反映该任务独特挑战的强大基线显得尤为关键。最近，基于 vision transformer (ViT) 的基础模型在医学影像分类等任务中逐渐流行起来，这些模型通常是在数百万张组织病理学切片上通过自监督学习训练的。这些模型通常通过线性探测或参数高效的微调技术如 Low Rank Adaptation (LoRA) (Hu et al., 2022) 来使用，在线性探测中，会在冻结的主干网络之上训练一个轻量级分类器。但是缺乏针对这些方法在非典型与正常有丝分裂分类中的综合研究。

本工作的贡献是两方面的：首先，我们提供了三个数据集，围绕识别 AMFs 展开；其次，我们提供了一个全面的基准测试，比较了广泛使用的深度学习架构、带有线性探测的基础模型以及针对非典型与正常有丝分裂分类任务进行了 LoRA 微调的基础模型。

2. 数据集

我们在本文中介绍并使用了三个不同的数据集——AMi-Br、AtNorM-Br 和 AtNorM-MD，如下所述。这项工作是我们团队之前会议贡献 (Bertram et al., 2025) 的扩展，其中已经简要介绍了这些数据集中的第一个（AMi-Br）。

作为该集合中规模最大的数据集，我们在所有评估中使用来自 AMI-Br 数据集的数据进行训练。这使我们可以将另外两个新引入的数据集用作独立的保留集。已经表明，数据分布变化（领域偏移）显著阻碍了深度学习模型的表现 (Stacke et al., 2020; Aubreville et al., 2021)。通过这些数据集，我们可以预期不同程度的领域偏移，因为它们来源于不同的来源（AtNorm-Br）或甚至是不同类型的肿瘤（AtNorm-MD），这使我们能够研究泛化能力。该数据集的统计数据，包括每个数据集中非典型和正常有丝分裂图的数量、注释类型、来源数据集、涉及物种等，汇总于表 1 中。

表 1: 数据集统计包括样本数量、类别平衡性、标注详情和领域覆盖率。

属性	AMi-Br	在诺姆-MD	在诺姆-布雷
Total MFs	3,720	2,107	746
Atypical	832	219	128
Normal	2,888	1,888	618
Atypical Rate (%)	22.37%	10.39%	17.16%
Annotation Type	3-expert vote	5-expert vote	Single expert
Expert Agreement	78.2%	69.6%	–
Source Dataset(s)	TUPAC16, MIDOG21	MIDOG++	TCGA (BRCA)
Species	Human	Human + Canine	Human

2.1 AMi-Br

第一个数据集，AMi-Br (Bertram et al., 2025) 包含了总共 3,720 个来自人类乳腺癌的 MFs，其中 1,999 个 MFs 来自于 TUPAC16 (Veta et al., 2019) 的替代标签集 (Bertram et al., 2020)，而 1,721 个 MFs 则来自 MIDOG21 训练数据集 (Aubreville et al., 2023b)。MFs 被三位专家病理学家分类为两类之一——非典型或正常。根据三位专家之间的多数投票结果，发现有 832 个 MFs 是非典型的，而剩余的 2,888 个被归类为正常的 MFs。我们发现在所有病例中，共有 2908（78.17%）个对象得到了三位专家的一致意见。

2.2 在诺姆-MD

第二个数据集，Atypical and Normal Mitosis (AtNorM)-多领域（MD）扩展到人类乳腺癌之外，涵盖六个领域，包括人类和犬类肿瘤。这些包括犬类肺癌、犬类淋巴瘤、犬类皮肤肥大细胞肿瘤、人类神经内分泌肿瘤、犬类软组织肉瘤和人类黑色素瘤。它来源于公开可用的 MIDOG++数据集中的所有领域，除了人类乳腺癌领域 (Aubreville et al., 2023c)。它包含 2,107 个有丝分裂体，其中 219 个（10.4%）为非典型。标记是由五位病理学专家进行多数投票，其中三位是获得认证的。我们发现，在 1,466 个病例（69.58%）中，所有五位专家对每个对象达成完全一致。

2.3 在范式-边界

第三个数据集，AtNorM-Br，在本研究的范围内也公开提供，包含来自 The Cancer Genome Atlas (TCGA) (Lingle et al., 2016) 乳腺癌 (BRCA) 队列中 179 名患者的 746 个 MF 实例。TCGA 包含来自各种来源且质量部分混杂的图像，因此也是评估泛化能力的重要资产。该数据集由一位具有高经验的 AMF 分类专家进行标注，根据标注结果，发现了 128 个 MFs 是非典型的，剩下的 618 个被分类为正常，从而得出异常率（AMF）为 17.16%。

表 2: 本文使用的基础模型概述

模型	预训练算法	训练数据集大小		模型类型	大小（参数）
		WSIs	瓷砖		
UNI	DINOv2	100K	100M	ViT-L/16	307M
UNI2-h	DINOv2	350K	200M	ViT-H/14	681M
Virchow	DINOv2	1.5M	2B	ViT-H/14	632M
Virchow2	DINOv2	3.4M	1.7B	ViT-H/14	632M
Prov-Gigapath	DINOv2	170K	1.3B	ViT-g/16	1.1B
H-Optimus-0	DINOv2	500K	-	ViT-g/14	1.1B
H-Optimus-1	DINOv2	1M	2B	ViT-g/14	1.13B

3. 方法

3.1 端到端训练的深度学习模型

为了建立非典型与正常有丝分裂分类的稳健基线，我们评估了三种广泛使用的深度学习架构——EfficientNetV2 (Tan and Le, 2021), ViT (Dosovitskiy et al., 2020)，和 Swin Transformer (Liu et al., 2021)。所有这些模型都在各种医学成像数据集和任务中展示了强大的性能。EfficientNetV2 是一种众所周知的 convolutional neural network (CNN)，以其在各类分类任务中的强大性能和效率而著称。它作为传统基于 CNN 方法的代表。相比之下，ViT 代表了从卷积架构到基于变压器模型的范式转变。一个 ViT 将图像处理为补丁序列，并利用自注意力机制来建模输入之间的全局关系。这种全局感受野使 ViT 能够捕捉长距离依赖，这在复杂的组织病理学模式中特别有用。Swin Transformer 通过引入分层架构和移位窗口进一步推进了基于变压器的方法。这种设计使得模型能够在局部计算注意力的同时逐渐构建出全局表示，实现了计算效率与捕获局部细节和全局结构能力之间的平衡。Swin Transformers 在密集预测任务和高分辨率图像分析中特别有效，使其非常适合有丝分裂图形分类 (Liu et al., 2021)。此外，所选的模型架构也被选择为相似大小。每个模型都以端到端的方式进行训练，并在一个随机的、基于案例分割的 AMi-Br 数据集上进行了评估：该数据集占了数据集中样本的 ≈22%（未用于训练或验证的样本），以及整个 AtNorM-MD 和 AtNorM-Br 数据集。

3.2 基础模型

基础模型是在大量多样化的数据集上训练的模型，通常使用自我监督等无监督技术进行训练。训练的目标是培养能够很好地泛化并易于适应下游任务的功能提取器。在计算病理学领域，我们已经看到公开可用的模型有了非常显著的增长，这需要对 AMF 分类任务进行分析。我们将八种最先进的模型进行了比较 – UNI (Chen et al., 2024), UNI2-h (Chen et al., 2024; MahmoodLab, 2025), Virchow (Vorontsov

et al., 2024)，Virchow2 (Zimmermann et al., 2024)，Prov-Gigapath (Xu et al., 2024)，H-Optimus-0 (Saillard et al., 2024)，H-Optimus-1 (Bioptimus, 2025) 和 H0-mini (Filiot et al., 2025)。每个模型的详细信息概述见表 2。所有选定的基础模型都是基于 ViT，并已在包含数十万到数百万个 whole slide images (WSIs) 的数据集上进行了预训练。

3.2.1 基础模型的线性探测

我们采用线性探测策略，其中特征提取器（即基础模型）用于抽取嵌入。之后，基础模型保持冻结状态，并在抽取的特征之上训练一个线性分类器。这一过程有助于我们评估基础模型表示区分非典型与正常有丝分裂的能力。

3.2.2 LoRA 微调

参数高效微调 (PEFT) 方法 (Houlsby et al., 2019) 是一种模型调整策略，专注于仅对模型参数进行最小的修改来进行微调。传统的微调会更新所有模型参数，对于大型模型而言这在计算上非常昂贵。LoRA (Hu et al., 2022)，作为一种 PEFT 方法，通过冻结原始权重并在特定层引入小型可训练的低秩矩阵来解决这个问题，通常是在自注意力块中的查询、键和值权重矩阵内。只有这些矩阵会在训练过程中被更新，大大减少了内存和计算需求，同时保持了与全量微调相当的性能。我们将基于 LoRA 的微调应用于之前通过线性探测评估过的同一组模型——UNI、UNI2-h、Virchow、Virchow2、Prov-Gigapath、H-Optimus-0 和 H-Optimus-1，以进行直接比较。

4. 实验设置

我们为上述情况设置了详细的实验。二分类实验中的图像预处理包括将所有切片调整为 224×224 像素。训练增强包括随机水平翻转、旋转、颜色抖动和随机裁剪并缩放，随后使用 ImageNet 统计数据进行归一化。验证图像仅进行了缩放和归一化。每个折叠的最佳检查点是根据验证平衡准确率选择的。所有实验均采用 5 折交叉验证进行，其中数据分层基于幻灯片（患者），以确保来自同一幻灯片的所有切片只出现在一个拆分中（训练或验证），从而防止数据泄漏。为缓解类别不平衡问题，我们在训练期间使用了加权随机采样，样本权重与类别频率成反比。所有模型均使用 Adam 优化器进行训练，并带有 L2 正则化（学习率：1e-4，权重衰减：1e-5）和交叉熵损失。我们基于验证平衡准确率应用了早停策略，耐心值为 15 个周期，并在平台期采用了学习率调度（因子：0.5，耐心：3，最小 LR：1e-7）。每个模型最多训练 100 个周期，批量大小为 8。其余的案例特定细节在以下子部分中描述。

4.1 端到端训练的深度学习模型

我们将所有三个以端到端方式训练的模型（即 EfficientNetV2、ViT 和 Swin Transformer）用 ImageNet 预训练权重进行了初始化，并将原始分类头替换为一个由单一线性层组成的自定义二元分类器。具体来说，我们使用了来自时间管理模块库的 `vit_large_patch16_224` 模型，这是一个输入分辨率为 224×224、块大小为 16×16 且包含 8660 万参数的 ViT。所使用的 Swin Transformer 变体是 `swin_base_patch4_window7_224`，其特征为 4×4 的块大小和 7×7 的移位注意力窗口，共计包含 8780 万个参数。对于 EfficientNetV2，我们使用了来自及时的 `efficientnetv2_m`，它大约有 5410 万个参数。此协议同样适用于所有三个模型，以公平比较它们在非典型有丝分裂分类任务上的性能。

4.2 基于基础模型的线性探测

我们使用标准的线性探测方法在 AMi-Br 数据集上对七个基础模型进行了全面评估，该数据集用于非典型与正常有丝分裂分类。对于每个模型，我们都从组织病理学图像中提取高维特征嵌入，并训练单个线性分类层，同时保持预训练的特征提取器不变。这种方法在所有六个基础模型中一致应用，以确保在其非典型分类任务上的特征表示能力进行公平比较。

4.3 LoRA 基础模型的微调

对于基于 LoRA 的基础模型微调，每个模型均使用公开的预训练权重初始化，并通过秩为 8、缩放因子为 16 以及 dropout 率为 0.3 的 LoRA 进行调整，应用于 transformer 注意力和 MLP 层。分类头重新初始化并与 LoRA 模块联合训练，而其余骨干部分保持冻结。这一标准化协议确保了在 AMi-Br 基准上对各种基础模型使用基于 LoRA 的适应性评估公平且严格。

5. 结果

在 AMi-Br 数据集上，表现最佳的模型是进行 LoRA 微调的 Virchow2，实现了平均平衡准确率为 0.8135 和 AUROC 为 0.9026，超越了所有其他模型。其他如 UNI（平均平衡准确率 0.7952）和 Virchow（平衡准确率 0.7878）等进行 LoRA 微调的模型也表现出色，显著优于各自线性探测版本的表现。在端到端训练的基线模型中，ViT 达到了最高的平衡准确率（0.7552）和 AUROC（0.8634），超越了 EfficientNetV2

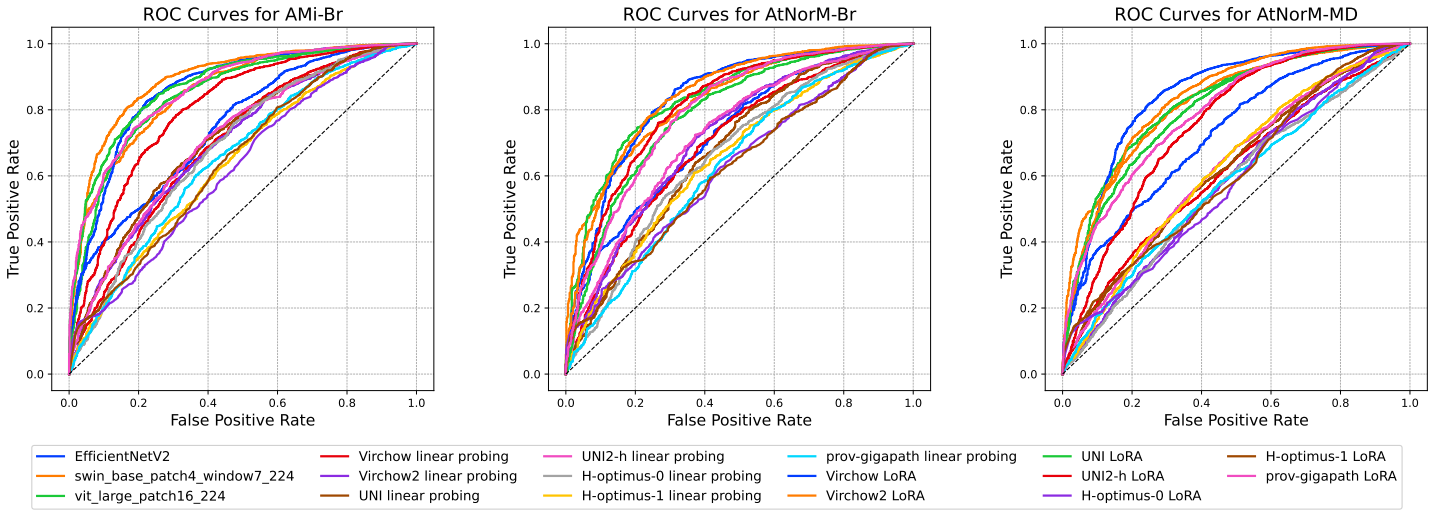

图 2: Receiver operating characteristic (ROC) 曲线用于基线、基础模型的线性探测和基础模型的 LoRA 微调，涵盖三个数据集。

和 Swin Transformer。仅进行线性探测的基础模型落后于其他模型，其平衡准确率为大约 0.59-0.64。

在 AtNorM-Br 数据集上，趋势相似。Virchow 通过 LoRA 微调达到了最高的平衡精度 0.7696，而 Virchow2 通过 LoRA 微调达到了最高的 AUROC 0.8579，其次是通过 LoRA 微调的 Virchow2（平衡精度：0.7632）。其他经过 LoRA 微调的模型，如 UNI2-h（0.7301）和 UNI（0.7183），也显示出比它们的线性探测变体有持续改进。在基线中，Vision Transformer 保持了强劲的表现（平衡精度：0.7570，AUROC：0.8478），再次优于 EfficientNetV2 和 Swin Transformer。没有经过 LoRA 适应的基础模型表现一般，平衡精度在 0.60 到 0.66 之间。

AtNorM-MD 数据集代表了一种显著的分布变化，模型在该数据集上的性能普遍下降。预训练了 LoRA 的 Virchow 表现最佳，其平衡准确率为 0.7705，AUROC 值为 0.8641。其他表现优秀的 LoRA 微调模型包括 Virchow2（平衡准确率 0.7424）和 UNI（平衡准确率 0.7069），这些模型的性能均显著优于其线性探针基线模型。在此设置下，最佳基准仍然是 Vision Transformer，实现了 0.7439 的平衡准确率和 0.8396 的 AUROC 值。相比之下，没有 LoRA 迁移的预训练基础模型性能较不稳定。总体而言，在领域变化的情况下，LoRA 微调始终能持续提升泛化性能。

在所有数据集中，两个模型 H-Optimus-0 和 H-Optimus-1，均使用 LoRA 进行了微调，表现一直不佳，甚至低于其线性探测对应的平衡准确率。

6. 讨论

我们的评估确认所有基础模型在 AMF 识别任务中表现出一定程度的泛化能力，但它们的有效性程度差异很大。应用 LoRA 微调揭示了各模型之间的性能差异尤为显著。

H-Optimus-0 和 H-Optimus-1 的表现不佳可能是由于分辨率不匹配造成的，因为该模型期望输入图像的分辨率为 0.5 微米/像素，而我们的数据集中的图像是以 0.25 微米/像素的分辨率提供的。相比之下，Virchow2 表现更优，这可能得益于使用 40 倍物镜训练高分辨率图像，使其与我们数据集的特点更加兼容。

我们还注意到基础模型（包括线性探测和 LoRAs-微调）在两个外部测试数据集中的性能显著下降，这一点反映在平衡准确率和 AUROC 分数上。我们认为这是由于数据集之间的领域变化所致。对于 AtNorM-Br 数据集，观察到的性能可能部分也源于鉴别 AMFs 时个体标准的不同，当多个评估者参与时这种影响会趋于平缓。由于只有单一标注员进行了 AtNorM-Br 的数据标记过程，这可能会引入潜在的系统性标签偏差。此外，也不能排除图像分布变化的可能性，因为图像是从不同的来源（实验室、扫描仪等）获取的，使得数据集更加异质化。对于有五个评估者参与的 AtNorM-MD 数据集，出现系统性标签偏差的可能性较小，然而，包含多个领域的乳腺癌以外的数据引入了更为显著的图像领域变化，涵盖了对模型适应构成重大挑战的各种组织类型。

我们方法的一个重要限制在于仅在 AMi-Br 数据集上进行训练，该数据集中只有相对较少的人类乳腺癌病例，这限制了模型学习不同组织类型、肿瘤、机构、染色协议和扫描仪的 AMF 表示能力。我们的结果强调，自动 AMF 识

表 3: 不同模型在三个测试数据集上的性能。值以平均值 ± 标准差报告。

模型	AMi-Br		在范式-边界		在诺姆-_md	
	平衡精度	AUROC	平衡准确率	AUROC	平衡准确率	AUROC
EfficientNetV2	0.6488 ± 0.1265	0.7124 ± 0.1608	0.6457 ± 0.1055	0.7055 ± 0.1250	0.6034 ± 0.1166	0.6721 ± 0.1279
Vision Transformer	0.7552 ± 0.0191	0.8634 ± 0.0104	0.7570 ± 0.0171	0.8478 ± 0.0189	0.7439 ± 0.0149	0.8396 ± 0.0213
Swin Transformer	0.7491 ± 0.1061	0.8451 ± 0.0996	0.7213 ± 0.1116	0.8272 ± 0.0879	0.7226 ± 0.1048	0.8069 ± 0.1213
UNI	0.6477 ± 0.0108	0.7171 ± 0.0080	0.6217 ± 0.0015	0.6774 ± 0.0136	0.5715 ± 0.0153	0.6241 ± 0.0222
UNI2-h	0.6472 ± 0.0138	0.7087 ± 0.0107	0.6612 ± 0.0169	0.7296 ± 0.0115	0.5919 ± 0.0162	0.6297 ± 0.0173
Virchow	0.6452 ± 0.0139	0.7010 ± 0.0077	0.6441 ± 0.0212	0.7037 ± 0.0107	0.5758 ± 0.0172	0.6175 ± 0.0227
Virchow2	0.6403 ± 0.0079	0.7040 ± 0.0052	0.6667 ± 0.0133	0.7201 ± 0.0060	0.5652 ± 0.0064	0.5878 ± 0.0211
Prov-Gigapath	0.6091 ± 0.0140	0.6569 ± 0.0132	0.6005 ± 0.0091	0.6323 ± 0.0125	0.5601 ± 0.0153	0.5818 ± 0.0129
H-Optimus-0	0.6372 ± 0.0164	0.6926 ± 0.0062	0.6122 ± 0.0185	0.6903 ± 0.0217	0.5603 ± 0.0400	0.5935 ± 0.0423
H-Optimus-1	0.5918 ± 0.0165	0.6494 ± 0.0108	0.6139 ± 0.0195	0.6696 ± 0.0193	0.5915 ± 0.0308	0.6436 ± 0.0499
UNI (LoRA)	0.7952 ± 0.0092	0.8839 ± 0.0059	0.7183 ± 0.0226	0.7979 ± 0.0142	0.7069 ± 0.0278	0.8222 ± 0.0224
UNI2-h (LoRA)	0.7138 ± 0.0121	0.8153 ± 0.0211	0.7301 ± 0.0152	0.8228 ± 0.0143	0.6914 ± 0.0321	0.7616 ± 0.0415
Virchow (LoRA)	0.7878 ± 0.0250	0.8891 ± 0.0150	$\mathbf{0.7696} \pm 0.0198$	0.8540 ± 0.0200	$\mathbf{0.7705} \pm 0.0287$	$\mathbf{0.8641} \pm 0.0247$
Virchow2 (LoRA)	$\mathbf{0.8135} \pm 0.0145$	$\mathbf{0.9026} \pm 0.0051$	0.7632 ± 0.0190	$\mathbf{0.8579} \pm 0.0117$	0.7424 ± 0.0305	0.8503 ± 0.0171
Prov-Gigapath (LoRA)	0.7602 ± 0.0113	0.8682 ± 0.0122	0.7263 ± 0.0296	0.8077 ± 0.0184	0.7007 ± 0.0228	0.8073 ± 0.0259
H-Optimus-0 (LoRA)	0.5888 ± 0.0766	0.6159 ± 0.1276	0.5901 ± 0.0827	0.6224 ± 0.1220	0.5406 ± 0.0619	0.5826 ± 0.1263
H-Optimus-1 (LoRA)	0.5908 ± 0.0909	0.6362 ± 0.1143	0.5699 ± 0.0752	0.6352 ± 0.1091	0.5617 ± 0.0874	0.6019 ± 0.1312

别仍然是计算病理学中的一个高度挑战性问题，需要进一步的方法改进。本研究提供的新的多领域数据集有助于推进这一目标。

Acknowledgments

CAB、VW 和 CK 感谢奥地利研究基金（FWF，项目编号：I 6555）的支持。SB、TC、RK 和 MA 感谢德国研究基金会（DFG，德国研究基金会，项目编号：520330054）的支持。KB 感谢德国研究基金会（DFG）项目 460333672 CRC1540 EBM 的支持。

Ethical Standards

本工作在进行研究和撰写手稿时遵循了适当的伦理标准，并遵守了所有关于动物或人类受试者治疗的相关法律法规。

Conflicts of Interest

我们声明不存在利益冲突。

Data availability

本研究中使用的所有数据均为公开数据，可在以下 github 仓库找到：`https://github.com/DeepMicroscopy/AMi-Br_Benchmark`。

References

Marc Aubreville, Christof Bertram, Mitko Veta, Robert Klopfleisch, Nikolas Stathonikos, Katharina Breininger, Natalie ter Hoeve, Francesco Ciompi, and Andreas Maier. Quantifying the scanner-induced domain gap in mitosis detection. *Medical Imaging with Deep Learning*, 2021. URL `https://2021.midl.io/papers/i6`.

Marc Aubreville, Jonathan Ganz, Jonas Ammeling, Taryn A Donovan, Rutger HJ Fick, Katharina Breininger, and Christof A Bertram. Deep learning-based subtyping of atypical and normal mitoses using a hierarchical anchor-free object detector. In *BVM Workshop*, pages 189–195. Springer, 2023a.

Marc Aubreville, Nikolas Stathonikos, Christof A Bertram, Robert Klopfleisch, Natalie Ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A Donovan, Andreas Maier, et al. Mitosis domain generalization in histopathology images—the MIDOG challenge. *Medical Image Analysis*, 84:102699, 2023b.

Marc Aubreville, Frauke Wilm, Nikolas Stathonikos, Katharina Breininger, Taryn A Donovan, Samir Jabari, Mitko Veta, Jonathan Ganz, Jonas Ammeling, Paul J van Diest, et al. A comprehensive multi-domain dataset for mitotic figure detection. *Scientific data*, 10(1):484, 2023c.

Marc Aubreville, Nikolas Stathonikos, Taryn A Donovan,

Robert Klopfleisch, Jonas Ammeling, Jonathan Ganz, Frauke Wilm, Mitko Veta, Samir Jabari, Markus Eckstein, et al. Domain generalization across tumor types, laboratories, and species—insights from the 2022 edition of the mitosis domain generalization challenge. *Medical Image Analysis*, 94:103155, 2024.

Christof A Bertram, Mitko Veta, Christian Marzahl, Nikolas Stathonikos, Andreas Maier, Robert Klopfleisch, and Marc Aubreville. Are pathologist-defined labels reproducible? comparison of the tupac16 mitotic figure dataset with an alternative set of labels. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pages 204–213. Springer, 2020.

Christof A Bertram, Alexander Bartel, Taryn A Donovan, and Matti Kiupel. Atypical mitotic figures are prognostically meaningful for canine cutaneous mast cell tumors. *Veterinary Sciences*, 11(1):5, 2023.

Christof A Bertram, Viktoria Weiss, Taryn A Donovan, Sweta Banerjee, Thomas Conrad, Jonas Ammeling, Robert Klopfleisch, Christopher Kaltenecker, and Marc Aubreville. Histologic dataset of normal and atypical mitotic figures on human breast cancer (ami-br). In *BVM Workshop*, pages 113–118. Springer, 2025.

Bioptimus. H-optimus-1, 2025. URL `https://huggingface.co/bioptimus/H-optimus-1`.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

Taryn A Donovan, Frances M Moore, Christof A Bertram, Richard Luong, Pompei Bolfa, Robert Klopfleisch, Harold Tvedten, Elisa N Salas, Derick B Whitley, Marc Aubreville, et al. Mitotic figures—normal, atypical, and imposters: A guide to identification. *Veterinary pathology*, 58(2): 243–257, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Rutger RH Fick, Christof Bertram, and Marc Aubreville. Improving cnn-based mitosis detection through rescanning annotated glass slides and atypical mitosis subtyping. In *Medical Imaging with Deep Learning*, 2024.

Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology, 2025. URL `https://arxiv.org/abs/2501.16239`.

Patrick L Fitzgibbons and James L Connolly. Protocol for the examination of resection specimens from patients with invasive carcinoma of the breast. *CAP guidelines*, 4.8.1.0, 2023. URL `https://www.cap.org/cancerprotocols`.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Mostafa Jahanifar, Muhammad Dawood, Neda Zamanitajeddin, Adam Shephard, Brinder Singh Chohan, Christof A Bertram, Noorul Wahab, Mark Eastwood, Marc Aubreville, Shan E Ahmed Raza, et al. Pan-cancer profiling of mitotic topology & mitotic errors: Insights into prognosis, genomic alterations, and immune landscape. *medRxiv*, pages 2025–06, 2025.

Yuesheng Jin, Ylva Stewénius, David Lindgren, Attila Frigyesi, Olga Calcagnile, Tord Jonson, Anna Edqvist, Nina Larsson, Lena Maria Lundberg, Gunilla Chebil, et al. Distinct mitotic segregation errors mediate chromosomal instability in

aggressive urothelial cancers. *Clinical cancer research*, 13 (6):1703–1712, 2007.

Beata Kalatova, Renata Jesenska, Daniel Hlinka, and Marek Dudas. Tripolar mitosis in human cells and embryos: occurrence, pathophysiology and medical implications. *Acta histochemica*, 117(1):111–125, 2015.

M Kiupel, JD Webster, KL Bailey, S Best, J DeLay, CJ Detrisac, SD Fitzgerald, D Gamble, PE Ginn, MH Goldschmidt, et al. Proposal of a 2-tier histologic grading system for canine cutaneous mast cell tumors to more accurately predict biological behavior. *Vet. Pathol.*, 48(1):147–155, 2011. .

Ayat Lashen, Michael S Toss, Mansour Alsaleem, Andrew R Green, Nigel P Mongan, and Emad Rakha. The characteristics and clinical significance of atypical mitosis in breast cancer. *Modern Pathology*, 35(10):1341–1348, 2022.

Wilma Lingle, Bradley J Erickson, Margarita L Zuley, Rose Jarosz, Ermelinda Bonaccio, Joe Filippini, Jose M Net, Len Levi, Elizabeth A Morris, Gloria G Figler, et al. The cancer genome atlas breast invasive carcinoma collection (tcga-brca). *(No Title)*, 2016.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

David N Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, 131(6):803–820, 2016. .

MahmoodLab. Mahmoodlab/uni2-h. `https://huggingface.co/MahmoodLab/UNI2-h`, 2025. Accessed: 2025-06-26.

Yoko Matsuda, Hisashi Yoshimura, Toshiyuki Ishiwata, Hiroki Sumiyoshi, Akira Matsushita, Yoshiharu Nakamura, Junko Aida, Eiji Uchida, Kaiyo Takubo, and Tomio Arai. Mitotic index and multipolar mitosis in routine histologic sections as prognostic markers of pancreatic cancers: a clinicopathological study. *Pancreatology*, 16(1):127–132, 2016.

Laura Medri, Annalisa Volpi, Oriana Nanni, Anna Maria Vecci, Annita Mangia, Francesco Schittulli, Franco Padovani, Donata Casadei Giunchi, Alfredo Vito, Dino Amadori, et al. Prognostic relevance of mitotic activity in patients with node-negative breast cancer. *Modern pathology*, 16(11): 1067–1075, 2003.

John S Meyer, Consuelo Alvarez, Clara Milikowski, Neal Olson, Irma Russo, Jose Russo, Andrew Glass, Barbara A Zehnbauer, Karen Lister, and Reza Parwaresch. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: Reproducibility of grade and advantages of proliferation index. *Modern Pathology*, 18 (8):1067–1078, August 2005.

John S Meyer, Eric Cosatto, and Hans Peter Graf. Mitotic index of invasive breast carcinoma. Achieving clinically meaningful precision and evaluating tertial cutoffs. *Archives of pathology & laboratory medicine*, 133(11):1826–1833, November 2009.

Ryuji Ohashi, Shigeki Namimatsu, Takashi Sakatani, Zenya Naito, Hiroyuki Takei, and Akira Shimizu. Prognostic utility of atypical mitoses in patients with breast cancer: A comparative study with ki67 and phosphohistone h3. *Journal of surgical oncology*, 118(3):557–567, 2018.

Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. URL `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

Mitko Veta, Yujing J Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A Shah, Dayong Wang, Mikael Rousson, et al.

Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical image analysis*, 54: 111–121, 2019.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30 (10):2924–2935, 2024.

Frauke Wilm, Christof A Bertram, Christian Marzahl, Alexander Bartel, Taryn A Donovan, Charles-Antoine Assenmacher, Kathrin Becker, Mark Bennett, Sarah Corner, Brieuc Cossic, et al. Influence of inter-annotator variability on automatic mitotic figure assessment. In *Bildverarbeitung für die Medizin 2021: Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021*, pages 241–246. Springer, 2021.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.