MELBA Journal Machine Learning for Biomedical Imaging

Benchmarking Deep Learning and Vision Foundation Models for Atypical vs. Normal Mitosis Classification with Cross-Dataset Evaluation

Sweta **Banerjee** [1], Viktoria **Weiss** [2], Taryn A. **Donovan** [3], Rutger H.J. **Fick** [4], Thomas **Conrad** [5], Jonas **Ammeling** [6], Nils **Porsche** [1], Robert **Klopfleisch** [5], Christopher C. **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, Berlin, Germany
6 Technische Hochschule Ingolstadt, Ingolstadt, Germany
7 Medical University of Vienna, Austria
8 Julius-Maximilians-Universität Würzburg, Würzburg, Germany

Abstract

Atypical mitosis marks a deviation in the cell division process that has been shown be an independent prognostic marker for tumor malignancy. However, atypical mitosis classification remains challenging due to low prevalence, at times subtle morphological differences from normal mitotic figures, low inter-rater agreement among pathologists, and class imbalance in datasets. Building on the Atypical Mitosis dataset for Breast Cancer (AMi-Br), this study presents a comprehensive benchmark comparing deep learning approaches for automated atypical mitotic figure (AMF) classification, including end-to-end fine-tuned deep learning models, foundation models with linear probing, and foundation models fine-tuned with low-rank adaptation (LoRA). For rigorous evaluation, we further introduce two new held-out AMF datasets - AtNorM-Br, a dataset of mitotic figures from the TCGA breast cancer cohort, and AtNorM-MD, a multi-domain dataset of mitotic figures from a subset of the MIDOG++ training set. We found average balanced accuracy values of up to 0.8135, 0.7788, and 0.7723 on the in-domain AMi-Br and the out-of-domain AtNorm-Br and AtNorM-MD datasets, respectively. Our work shows that atypical mitotic figure classification, while being a challenging problem, can be effectively addressed through the use of recent advances in transfer learning and model fine-tuning techniques. We make all code and data used in this paper available in this github repository: `https://github.com/DeepMicroscopy/AMi-Br_Benchmark`.

Keywords

Atypical Mitosis, Deep Learning, Foundation Models, Classification, Benchmarking, Histopathology, low-rank adaptation

Article informations

https://doi.org/10.59275/j.melba.2026-6c1g

Volume 2026, Received: 2025-07-15, Published 2026-03-12

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

Special issue: MELBA–BVM 2025 Special Issue
Guest editors: Andreas Maier, Thomas Deserno, Heinz Handels, Klaus Maier-Hein, Christoph Palm, Thomas Tolxdorff, Katharina Breininger

1. Introduction

Mitosis is the process where a cell replicates its genetic material (DNA) and then divides into two identical daughter cells. This process enables the proliferation of cells, supporting physiological tissue growth and facilitating the replacement of old, damaged, or worn-out cells. The dividing cell may be observed microscopically as a mitotic figure (MF) with specific morphologies resembling the highly regulated cell division phases. However, cell proliferation does not occur only under physiological conditions; it is also a hallmark of cancer, where cell division is dysregulated and increased. Subsequently, quantifying cell proliferation is essential for estimating the aggressiveness of many human and animal cancers, including human breast cancer (Fitzgibbons and

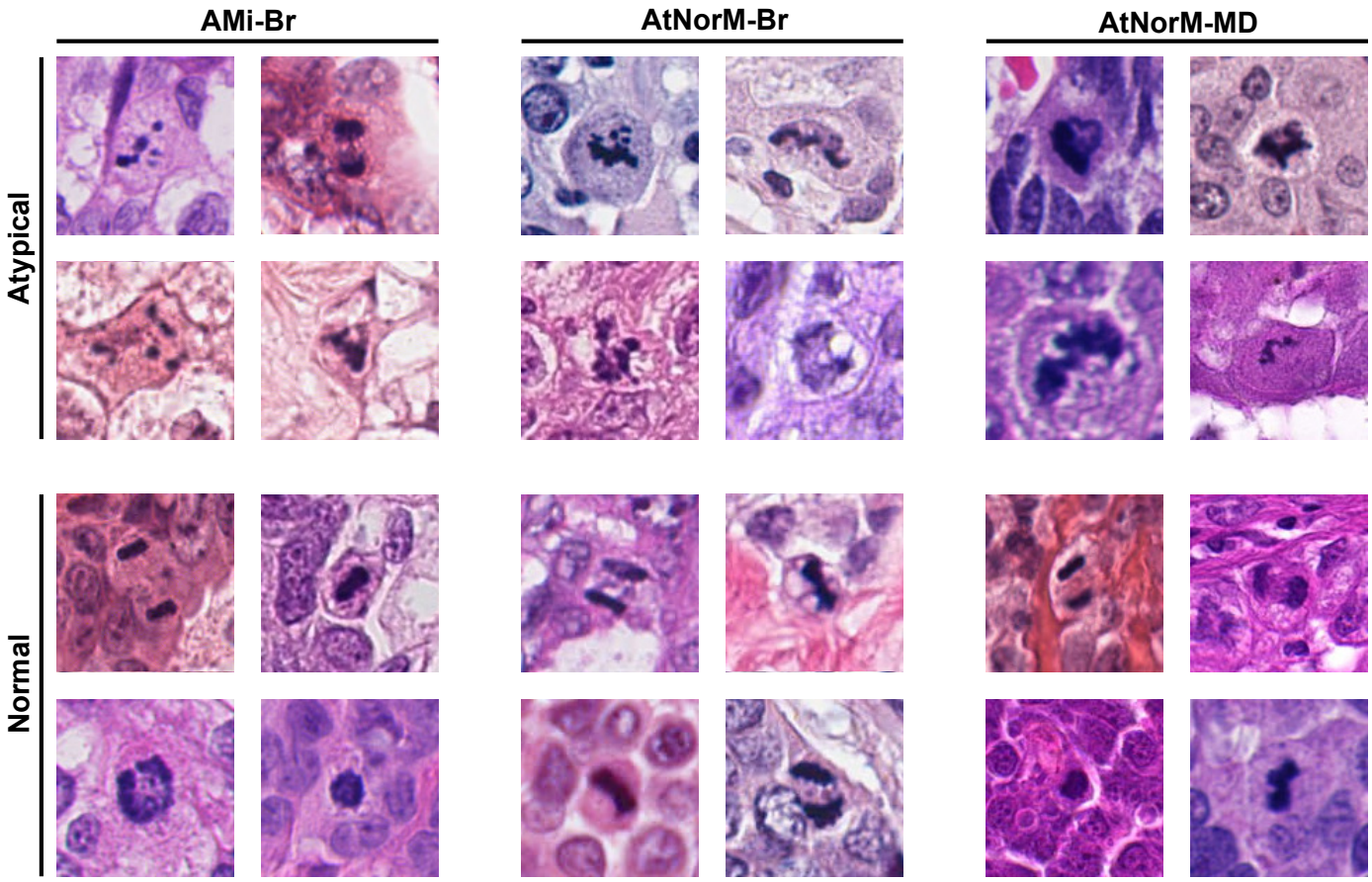

Figure 1: Image samples showing atypical and normal mitotic figures from all three datasets included in this study

Connolly, 2023; Kiupel et al., 2011; Louis et al., 2016). The routine method in the assessment of the growth rate is the mitotic count, which reflects the number of MFs within a specific tumor area, selected from a whole slide image (WSI). Elevated mitotic counts are associated with an increased risk of distant metastasis and higher risk of death (Medri et al., 2003).

In tumor samples, two variants of MFs can be found: normal and atypical. Normal MFs indicate the orderly process of cell division, designated by the prometaphase, metaphase, anaphase, and telophase, where chromosomes align and separate equally to produce two identical daughter cells (Donovan et al., 2021). An atypical mitotic figure (AMF), on the other hand, is a cell undergoing an abnormal cell division process characterized by errors in chromosome segregation or other morphological irregularities, which promote cancer progression (Donovan et al., 2021). Until recently, a correlation of the count of AMFs with shorter patient survival has been demonstrated in only few studies on breast cancer (Ohashi et al., 2018; Lashen et al., 2022) and selected other tumor types (Jin et al., 2007; Kalatova et al., 2015; Bertram et al., 2023; Matsuda et al., 2016). A very recent study by Jahanifar et al. (2025) has expanded these insights and the corresponding prognostic relevance of the rate of AMFs across a wide range of tumors in a large-scale evaluation, highlighting the importance of further investigations of this prognostic parameter.

Identification of MFs is traditionally done by pathologists, making it highly subjective and time intensive. This process is further complicated by inter-observer variability, low prevalence of MFs, and inconsistencies in slide quality (Meyer et al., 2005, 2009; Wilm et al., 2021). To address these challenges, deep learning-based MF detection has been extensively investigated (Aubreville et al., 2024; Veta et al., 2019; Aubreville et al., 2023b), demonstrating a promising outlook for improving both accuracy and efficiency and reducing the diagnostic workload for pathologists. Along similar lines, deep learning models offer potential for the automatic differentiation of AMFs against normal MFs. For instance, this could enable the quantification of the ratio of AMFs across entire whole slide images (Jahanifar et al., 2025), which is otherwise impractical to perform manually due to time constraints and the rarity of AMFs. Such quantitative assessments can support more objective and reproducible evaluations of tumor aggressiveness, potentially improving treatment planning in clinical settings.

At the time of this writing, there have been very few studies investigating deep learning for AMFs (Aubreville

et al., 2023a; Fick et al., 2024; Bertram et al., 2025). The automated classification of atypical vs. normal MFs is complicated by the substantial variation of morphologies, often with some mitotic figures displaying overlapping features of both normal and AMFs (Donovan et al., 2021; Bertram et al., 2023), and the low frequency of AMF, resulting in class imbalance. Given the prognostic relevance of differentiating normal vs. atypical MFs, standardized benchmarks, such as publicly available datasets, evaluation protocols, and baseline models, may serve as facilitators to push further developments in this field that tackle these challenges. In this context, foundation models offer a promising direction. These models, typically vision transformer (ViT)-based and pre-trained with self-supervised learning on millions of histopathology tiles, have shown strong performance across various medical imaging tasks, including classification (Campanella et al., 2025; Breen et al., 2025). When subsequently fine-tuned on comparatively small, task-specific labelled datasets, they are particularly effective in small data regimes. These models are typically employed either via linear probing, where a lightweight classifier is trained on top of a frozen backbone, or through parameter-efficient fine-tuning techniques like Low Rank Adaptation (LoRA) (Hu et al., 2022). However, comprehensive studies leveraging these techniques for atypical vs. normal mitosis classification are still lacking.

The contribution of this work is two-fold: First, we present three datasets, designed for the automated identification of AMFs to establish suitable datasets that reflect the unique challenges of this task. Second, we provide a comprehensive benchmark comparing end-to-end fine-tuned deep learning architectures, foundation models with linear probing, and foundation models fine-tuned with LoRA for the task of atypical versus normal mitotic figure classification.

2. Datasets

We present and use three distinct datasets in this paper - AMi-Br, AtNorM-Br, and AtNorM-MD as described below. This work is an extension of a previous conference contribution by our group (Bertram et al., 2025), which already introduced the first of those datasets (AMi-Br) in brief. Representative image samples of atypical and normal mitotic figures from all three datasets are shown in Figure 1.

Being the largest of the three datasets used in this study, we use data from the AMi-Br dataset for training in all evaluations. This allows us to use the other two newly introduced datasets as independent held-out sets. It has been shown that data distribution shifts (domain shifts) impede the performance of deep learning models considerably (Stacke et al., 2020; Aubreville et al., 2021). With these datasets, we expect varying degrees of domain shifts, and which allows us to investigate generalization performance. Only a part of the AMi-Br dataset is used for training the models for binary classification of atypical vs. normal mitoses, and the rest is used as another held-out test dataset. The dataset statistics, including total number of atypical and normal MFs in each, annotation type, source dataset(s), species involved etc. have been summarized in Table 1.

Table 1: Dataset statistics including sample counts, class balance, annotation details, and domain coverage.

Property	AMi-Br	AtNorM-Br	AtNorM-MD
Total MFs	3,720	746	2,107
Atypical	832	128	219
Normal	2,888	618	1,888
AMF Rate (%)	22.4%	17.2%	10.4%
Annotation Type	3-expert vote	Single expert	5-expert vote
Expert Agreement	78.2%	–	69.6%
Source Dataset(s)	TUPAC16, MIDOG21	TCGA (BRCA)	MIDOG++
Species	Human	Human	Human + Canine

2.1 AMi-Br

The first dataset, AMi-Br (Bertram et al., 2025) consists of a total of 3,720 MFs from human breast cancer, of which 1,999 MFs are from the TUPAC16 (Veta et al., 2019) alternative label set (Bertram et al., 2020) and 1,721 MFs are from the MIDOG21 training dataset (Aubreville et al., 2023b). The MFs were graded into one of two classes — atypical or normal — by three expert pathologists, two of whom were board-certified. According to the result of the majority vote between the three experts, 832 MFs were found to be atypical, while the remaining 2,888 were classified as normal MFs. We found total agreement per object by all three experts in 2908 (78.2%) of the cases.

2.2 AtNorM-Br

The third dataset, Atypical and Normal Mitosis (AtNorM)-Br, also made available publicly within the scope of this work, contains 746 MF instances from 179 patients from the breast cancer (BRCA) cohort of the The Cancer Genome Atlas (TCGA) (Lingle et al., 2016). TCGA contains images from various sources and with partially mixed quality, stemming from differences in staining protocols, scanning equipment, and tissue preparation across institutions, and is thus also a valuable asset for the assessment of generalization. This dataset was annotated by a single expert with high experience in AMF classification, according to which 128 MFs were found to be atypical and the remaining 618 were classified as normal, yielding an AMF rate of 17.16%.

Table 2: Overview of foundation models used in this paper

Model	Training Dataset Size		Model Type	Size (Params)
	WSIs	Tiles		
UNI	100K	100M	ViT-L/16	307M
UNI2-h	350K	200M	ViT-H/14	681M
Virchow	1.5M	2B	ViT-H/14	632M
Virchow2	3.4M	1.7B	ViT-H/14	632M
Prov-Gigapath	170K	1.3B	ViT-g/16	1.1B
H-Optimus-0	500K	-	ViT-g/14	1.1B
H-Optimus-1	1M	2B	ViT-g/14	1.13B

2.3 AtNorM-MD

The second dataset, AtNorM-MultiDomain (MD) extends beyond human breast cancer and covers six domains, spanning both human and canine tumors. These include canine lung cancer, canine lymphoma, canine cutaneous mast cell tumor, human neuroendocrine tumor, canine soft tissue sarcoma, and human melanoma. It was sampled randomly from all but the human breast cancer domains of the publicly available MIDOG++ dataset (Aubreville et al., 2023c). It comprises 2,107 MFs from 70 patients, with 219 (10.4%) being atypical. Labeling was performed as a majority vote of five pathology experts, three of whom were board-certified. We found total agreement per object by all five experts in 1,466 (69.6%) of the cases.

3. Methods

3.1 End-to-end fine-tuned Baseline Deep Learning Models

To establish robust baselines for atypical vs. normal mitoses classification, we evaluate three widely-used deep learning architectures – EfficientNetV2 (Tan and Le, 2021), ViT (Dosovitskiy et al., 2020), and Swin Transformer (Liu et al., 2021). All of these models have demonstrated strong performance across various medical imaging datasets and tasks. EfficientNetV2 is a well-established convolutional neural network (CNN) known for its strong performance and efficiency in various classification tasks. It serves as a representative of traditional CNN-based approaches. In contrast, the ViT represents a major shift from convolutional architectures to transformer-based models. A ViT processes images as sequences of patches and uses self-attention mechanisms to model global relationships across the input. This global receptive field allows the ViT to capture long-range dependencies, which can be particularly useful for complex histopathological patterns. The Swin Transformer further advances the transformer-based approach by introducing a hierarchical architecture with shifted windows. This architecture enables the model to compute attention locally while gradually building up to global representations, providing a balance between computational efficiency and the ability to capture both local details and global structure. Swin Transformers have been particularly effective in dense prediction tasks and high-resolution image analysis, making it well-suited for mitotic figure classification (Liu et al., 2021). The selected model architectures were chosen to be of similar size in parameters. Each model was fine-tuned in an end-to-end manner and evaluated on three held-out test datasets: a random, patient-stratified split of the AMi-Br dataset, accounting for ≈ 22 % of the samples in the dataset (samples not used in training or validation), and the entire AtNorM-MD and AtNorM-Br datasets.

3.2 Foundation Models

Foundation models are models that are trained on large and diverse corpora of data, often using unsupervised techniques like self supervision. The goal of foundation models is to train feature extractors that generalize well and are easily adaptable to down-stream tasks. In the field of computational pathology, we have seen a very strong incline towards publicly available foundation models, calling for an investigation on the task of AMF classification. We compare eight state-of-the-art models – 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), and H0-mini (Filiot et al., 2025). An overview of selected details for each model such as training dataset size, size in parameters and model type for training is provided in Table 2. All selected foundation models are ViT-based and have been pre-trained using the DINOv2 algorithm (Oquab et al., 2023) on datasets comprising of hundreds of thousands to multiple millions of WSIs.

3.2.1 Linear Probing of Foundation Models

We use a linear probing strategy, where the feature extractor (i.e., the foundation model) is used to extract the embeddings. After that, the foundation model is kept frozen, and a linear classifier is trained on top of the extracted features. This process helps us to assess the ability of the representations of the foundation models to classify atypical vs. normal MFs.

3.2.2 LoRA Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) methods (Houlsby et al., 2019) are model adaptation strategies which focus on fine-tuning models with minimal changes to their parameters. Traditional fine-tuning updates all model parameters, which becomes computationally expensive for large models. LoRA (Hu et al., 2022), a PEFT method, addresses this by freezing the original weights and introducing small, train-

able low-rank matrices into specific layers, typically within the query, key, and value weight matrices of self-attention blocks. Only these matrices are updated during training, greatly reducing memory and compute requirements while retaining performance comparable to full fine-tuning. We apply LoRA-based finetuning to the same set of models previously evaluated with linear probing – UNI, UNI2-h, Virchow, Virchow2, Prov-Gigapath, H-Optimus-0 and H-Optimus-1, for a direct comparison.

4. Experimental Setup

We set up detailed experiments for the above cases. All training was performed exclusively on the AMi-Br dataset. The image preprocessing for the binary classification experiments included resizing all patches, originally of the size 128×128, to 224×224 pixels. Training augmentations consisted of random horizontal flip, rotation, color jitter, and random resized crop, followed by normalization. For models initialized from ImageNet-pretrained weights and fine-tuned end-to-end, we used the standard ImageNet mean and standard deviation for normalization. For the computational pathology foundation models used in the linear probing and LoRA experiments, we instead followed the pre-processing recommended for each model, i.e., normalizing with the model-specific mean and standard deviation used during their pre-training, rather than assuming ImageNet statistics. Validation images were only resized and normalized. The best checkpoint for each fold was selected based on validation balanced accuracy. The balanced accuracy for binary classification is the average of sensitivity and specificity and can be mathematically stated as follows:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

The balanced accuracy used as the primary metric because of the class imbalance between the atypical and normal MFs in the dataset, with a fixed decision threshold of 0.5 on the predicted probabilities. All experiments were performed using 5-fold cross validation, where stratification of the data was based on the slide (patient) to ensure that all patches from a given slide appeared in only one split (train or validation), thus preventing data leakage. To address class imbalance during training, we used a weighted random sampler that assigns higher sampling weights to under-represented classes and lower weights to frequent classes, with each sample's weight set inversely to its class frequency. We used weighted random sampling because it was proven to give the best results in previous work (Bertram et al., 2025). All models were trained using the Adam optimizer with L2 regularization (learning rate: 5×10^{-5}, weight decay: 1×10^{-5}) and cross-entropy loss. We applied early stopping with a patience of 15 epochs based on validation balanced accuracy, and used learning rate scheduling on plateau (factor: 0.5, patience: 3, minimum LR: 1×10^{-7}). Each model was trained for up to 100 epochs with a batch size of 8. The rest of the case-specific details are described below in the subsections.

4.1 End-to-end fine-tuned Baseline Deep Learning Models

We initialized all three models (i.e., EfficientNetV2, ViT and Swin Transformer) that were fine-tuned in an end-to-end fashion, essentially doing a full finetuning of all the layers with ImageNet-pretrained weights, and the original classification head was replaced with a custom binary classifier consisting of a single linear layer. Specifically, we used the `vit_large_patch16_224` model from the `timm` library (Wightman, 2019), a ViT with 224×224 input resolution, 16×16 patch size, and 86.6 million parameters. The Swin Transformer variant used was `swin_base_patch4_window7_224` which features a 4×4 patch size and 7×7 shifted attention windows, comprising 87.8 million parameters. For EfficientNetV2, we employed `efficientnetv2_m` from `timm`, which has approximately 54.1 million parameters. This protocol was applied identically to all three models to enable fair comparison of their performance on the atypical mitosis classification task.

4.2 Linear Probing with Foundation Models

We conducted comprehensive evaluation of seven foundation models using a standardized linear probing method for atypical versus normal mitosis classification. For each model, we extracted high-dimensional feature embeddings from histopathology images and trained a single linear classification layer while keeping the pre-trained feature extractor frozen. This method was consistently applied across all six foundation models to ensure fair comparison of their feature representation capabilities on the atypical classification task.

4.3 LoRA Fine-tuning of Foundation Models

For the LoRA-based fine-tuning of the foundation models, each model was initialized with publicly available pre-trained weights and adapted using LoRA with rank 8, scaling factor 16, and dropout rate 0.3, applied to both the transformer attention layers (i.e., query, key, value, and output projections) and the MLP components (i.e., feedforward layers fc1 and fc2) of the transformer blocks. The classification head was re-initialized and trained jointly with LoRA modules, while the rest of the backbone remained frozen. This standardized protocol ensured fair and rigorous evaluation of LoRA-based adaptation across diverse foundation models on the AMi-Br benchmark.

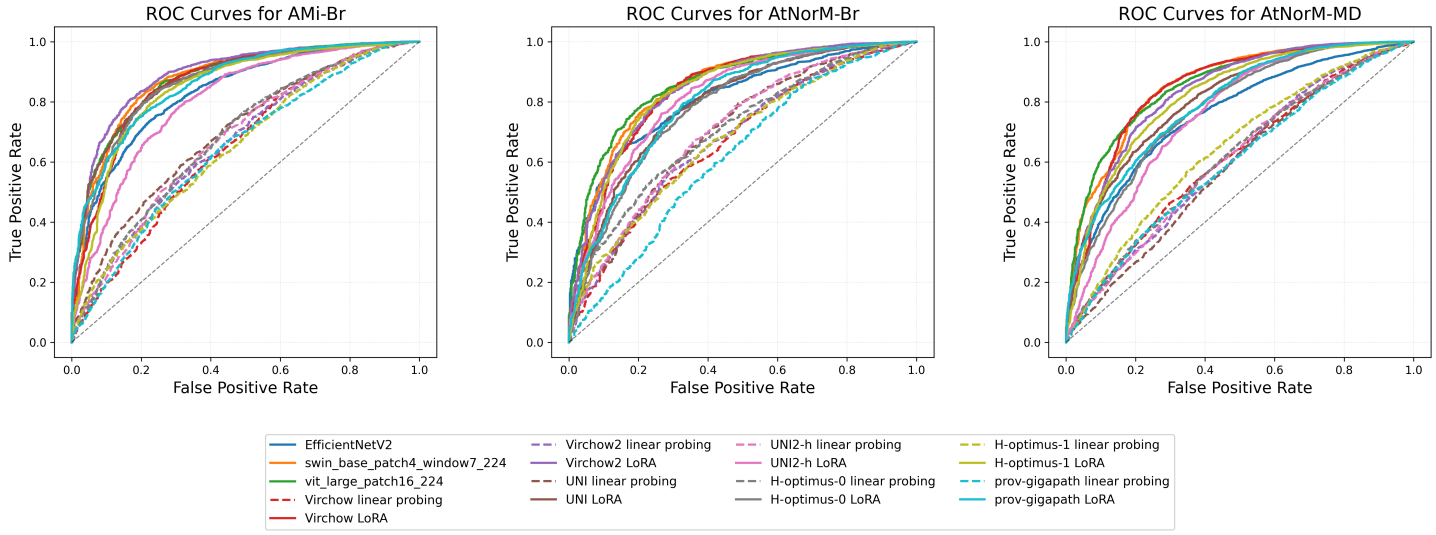

Figure 2: Receiver operating characteristic (ROC) curves for baseline, foundation models linear probing and LoRA finetuning of foundation models across three datasets.

5. Results

We now present the comparative performance of all models across the AMi-Br, AtNorM-Br, and AtNorM-MD datasets using balanced accuracy and AUROC as primary metrics. The full quantitative results are summarized in Table 3, where we report the mean and standard deviation on the hold-out datasets across five train-validation folds of the training part of AMi-Br. To complement the tabulated results, Figure 2 illustrates the corresponding ROC curves for each dataset, comparing baseline models, foundation models with linear probing, and those fine-tuned using LoRA.

On the AMi-Br dataset, the best-performing model in terms of balanced accuracy was Virchow2 with LoRA fine-tuning, achieving an average balanced accuracy of 0.8135 (sensitivity = 0.7511, specificity = 0.8760 of the atypical class). While the highest AUROC of 0.9029 was obtained by the Swin Transformer model, Virchow2 closely followed with an AUROC of 0.9026, making it the most balanced performer overall. Other LoRA-fine-tuned models such as UNI (0.7952 mean balanced accuracy), Virchow (0.7878 mean balanced accuracy) and H-Optimus-0 (0.7846 mean balanced accuracy) also demonstrated strong performance, significantly improving over their respective linear probing, as demonstrated by the pairwise Wilcoxon signed-rank tests for each test dataset (see Tab. 4). Among the end-to-end fine-tuned baseline models, the Swin Transformer achieved the highest balanced accuracy (0.8052) and AUROC (0.9029), outperforming both EfficientNetV2 and ViT. Foundation models with just linear probing lagged behind, with balanced accuracies ranging between 0.60-0.65.

On the AtNorM-Br dataset, the end-to-end fine-tuned ViT was the best performing model overall with a mean balanced accuracy of 0.7788 (sensitivity = 0.7109, specificity = 0.8466 of the atypical class) and AUROC 0.8710, while the Swin Transformer followed closely with a mean balanced accuracy of 0.7752 and an AUROC of 0.8683. Among the LoRA-fine-tuned foundation models, Virchow performed the best in terms of a mean balanced accuracy of 0.7696, followed by H-Optimus-1 (0.7658 mean balanced accuracy) and Virchow2 (0.7632 mean balanced accuracy) Other LoRA-fine-tuned models such as UNI2-h (0.7301 balanced accuracy), Prov-Gigapath (0.7263 balanced accuracy), UNI (0.7183 balanced accuracy) and H-Optimus-0 (0.7044 balanced accuracy) also showed consistent improvements over their linear probing variants. Foundation models with linear probing performed poorly, with balanced accuracies ranging between 0.59 and 0.65.

The AtNorM-MD dataset, which represents a distinct distribution shift, saw a general drop in performance across models. The end-to-end fine-tuned Swin Transformer performed the best, with a mean balanced accuracy of 0.7723 (sensitivity = 0.6548, specificity = 0.8897 of the atypical class) and an AUROC of 0.8806. Virchow pretrained with LoRA followed closely, with a balanced accuracy of 0.7705. Other LoRA-fine-tuned models also performed well, including Virchow2 (0.7424 balanced accuracy), H-Optimus-1 (0.7396 balanced accuracy), UNI (0.7069 balanced accuracy), Prov-Gigapath (0.7007 balanced accuracy), UNI2-h (0.6914 balanced accuracy) and H-Optimus-0 (0.6549 balanced accuracy), all of which substantially outperformed their linear probing counterparts. Among the end-to-end fine-tuned baselines, ViT followed closely after Swin Transformer, achieving 0.7494 balanced accuracy and 0.8720 AUROC. EfficientNetV2 achieved the least balanced accuracy of 0.6975 among the baselines. As with the other datasets, the foundation models without LoRA adaptation

Table 3: Classification performance of the evaluated models on the AMi-Br, AtNorM-Br, and AtNorM-MD test sets. For each dataset, balanced accuracy and AUROC are reported as mean ± standard deviation across cross-validation folds. All models were trained exclusively on the AMi-Br dataset, while evaluation was performed on the respective test sets. Within each column, bold values denote the best-performing model for that dataset and in the respective group of models.

Model	AMi-Br		AtNorM-Br		AtNorM-MD	
	Balanced Acc.	AUROC	Balanced Acc.	AUROC	Balanced Acc.	AUROC
EfficientNetV2 (Baseline)	0.7474 ± 0.0167	0.8315 ± 0.0168	0.7283 ± 0.0168	0.8110 ± 0.0084	0.6975 ± 0.0277	0.7604 ± 0.0217
Vision Transformer (Baseline)	0.7899 ± 0.0268	0.8872 ± 0.0107	$\mathbf{0.7788} \pm 0.0229$	$\mathbf{0.8710} \pm 0.0075$	0.7494 ± 0.0342	0.8720 ± 0.0117
Swin Transformer (Baseline)	$\mathbf{0.8052} \pm 0.0124$	$\mathbf{0.9029} \pm 0.0112$	0.7752 ± 0.0134	0.8683 ± 0.0113	$\mathbf{0.7723} \pm 0.0142$	$\mathbf{0.8806} \pm 0.0087$
UNI	$\mathbf{0.6339} \pm 0.0189$	$\mathbf{0.6949} \pm 0.0275$	0.6406 ± 0.0339	0.7028 ± 0.0310	0.5510 ± 0.0123	0.5892 ± 0.0118
UNI2-h	0.6220 ± 0.0095	0.6765 ± 0.0157	$\mathbf{0.6482} \pm 0.0214$	0.7059 ± 0.0230	0.5773 ± 0.0212	0.6119 ± 0.0357
Virchow	0.6029 ± 0.0132	0.6520 ± 0.0168	0.5969 ± 0.0155	0.6733 ± 0.0266	0.5623 ± 0.0235	0.6184 ± 0.0566
Virchow2	0.6089 ± 0.0192	0.6637 ± 0.0214	0.6236 ± 0.0296	0.6795 ± 0.0344	0.5694 ± 0.0173	0.6046 ± 0.0324
Prov-Gigapath	0.6097 ± 0.0177	0.6435 ± 0.0188	0.5893 ± 0.0074	0.6187 ± 0.0193	0.5558 ± 0.0125	0.6039 ± 0.0244
H-Optimus-0	0.6319 ± 0.0153	0.6801 ± 0.0181	0.6369 ± 0.0147	$\mathbf{0.7090} \pm 0.0163$	0.5821 ± 0.0195	0.6142 ± 0.0207
H-Optimus-1	0.5943 ± 0.0132	0.6518 ± 0.0152	0.6271 ± 0.0168	0.6777 ± 0.0275	$\mathbf{0.5966} \pm 0.0317$	$\mathbf{0.6462} \pm 0.0461$
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)	0.7846 ± 0.0169	0.8721 ± 0.0148	0.7044 ± 0.0383	0.7921 ± 0.0271	0.6549 ± 0.0281	0.7879 ± 0.0163
H-Optimus-1 (LoRA)	0.7762 ± 0.0110	0.8611 ± 0.0083	0.7658 ± 0.0298	0.8524 ± 0.0100	0.7396 ± 0.0202	0.8389 ± 0.0104

Table 4: *p*-values for group comparisons of balanced accuracies within each dataset.

Dataset	Linear Probing vs LoRA (Wilcoxon signed-rank)	Baseline vs Linear Probing (Mann–Whitney U)	Baseline vs LoRA (Mann–Whitney U)
AMi-Br	5.82×10^{-11}	2.91×10^{-8}	0.73
AtNorM-Br	1.16×10^{-10}	2.91×10^{-8}	0.06
AtNorM-MD	5.82×10^{-11}	2.91×10^{-8}	0.09

performed less robustly, with balanced accuracies ranging between 0.55 and 0.60.

6. Discussion

Our evaluation confirmed that all foundation models exhibited some degree of generalization to the AMF identification task, but the extent of their effectiveness varied considerably. The application of LoRA fine-tuning revealed particularly striking performance differences across models. Also, it is worthwhile to note that across nearly all datasets and performance metrics (balanced accuracy and AUROC), there consistently exists at least one end-to-end fine-tuned baseline model that outperforms both linear probing and LoRA-adapted foundation models.

We attribute this to not sufficiently discriminatory features extracted by the foundation models, of which many were not even trained on high magnification ($40\times$) patches (Zimmermann et al., 2024), leading to a considerable data distribution shift. Thus, the stronger adaptation of the model (by using rank-reduced adaptation through LoRA) provides considerable benefit for improving the extracted features, while using the full capacity of the model (full fine-tuning) yields even better results.

Most importantly, the primary scope of our study was not to identify a single best-performing model for the respective datasets for the task of AMF classification, but rather to compare different learning strategies for AMF classification - full fine-tuning of ImageNet-pretrained models, LoRA-based adaptation of foundation models, and linear probing of foundation models. To investigate this research question in a rigorous way, we performed statistical tests across all datasets to assess whether the performance differences between these learning strategies are statistically significant, and the resulting p-values are summarized in Table 4. These results show that the differences between linear probing and LoRA are statistically significant for all three datasets ($p \ll 0.001$), with LoRA consistently outperforming linear probing. Likewise, baseline models significantly outperform linear probing for all datasets ($p \ll 0.0001$).

We also note that the results of EfficientNet on the AMi-Br test dataset is better than reported in previous work (Bertram et al., 2025). This can be attributed to the fact that the latter used EfficientNet V2-S, while we use EfficientNet V2-M in our experiments.

We also notice a drop in performance for several settings, including both end-to-end fine-tuned models and foundation models (both linear probing and LoRAs-fine-tuned) in the two external test datasets, as reflected by the balanced accuracy and AUROC scores. We attribute this to the domain shift between the datasets. For the AtNorM-Br dataset, the observed performance may partially also be influenced by

the reliance on a single expert involved in the classification of AMFs, potentially introducing systematic label biases and limiting generalizability, an effect that evens out when multiple raters are involved. Moreover, an image distribution shift cannot be ruled out, as the images were retrieved from different sources (labs, scanners, etc.), making the dataset more heterogeneous. For the AtNorM-MD dataset, where five annotators were involved, a systematic label bias is rather unlikely, however, the inclusion of multiple domains beyond breast cancer introduces a more pronounced image domain shift, encompassing a wider range of tissue types that pose serious challenges for model adaptation.

A significant limitation of our approach lies in training exclusively on the AMi-Br dataset with a limited number of cases of only human breast cancer, restricting the models' ability to learn the full spectrum of AMF representations across different tissue types, tumors, institutions, staining protocols, and scanners. Our results emphasize that automated AMF identification remains a highly challenging problem in computational pathology, requiring further methodological advances. Our new, multi-domain datasets, provided by this work, help further this cause. Future work will address the integration of these classifier into a two-stage mitotic figure detection approach, and investigate the role of false positive / negative detections.

Author contributions

SB wrote the main manuscript and carried out the experiments. VW, TAD, TC, RK, and CAB served as experts in the dataset annotation. SB and JA carried out the statistical analysis. CAB, KB, and MA co-wrote the manuscript and guided the project. RHJF organized and compiled the AtNorM-Br dataset with the support of CAB as pathology expert. CK provided scientific expertise. NP created Figure 1. All authors reviewed the manuscript.

Acknowledgments

CAB, VW, and CK acknowledge the support from the Austrian Research Fund (FWF, project number: I 6555). SB, TC, RK, and MA acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number: 520330054). KB acknowledges support by the German Research Foundation (DFG) project 460333672 CRC1540 EBM. JA acknowledges support by the Bavarian State Ministry of the Sciences and the Arts (project FOKUS-TML). NP and MA acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number: 545049923)

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare no conflicts of interest.

Data availability

All data used in this study is public and can be found at `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`.

Jack Breen, Katie Allen, Kieran Zucker, Lucy Godson, Nicolas M Orsi, and Nishant Ravikumar. A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification. *NPJ Precision Oncology*, 9(1):33, 2025.

Gabriele Campanella, Shengjia Chen, Manbir Singh, Ruchika Verma, Silke Muehlstedt, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, et al. A clinical benchmark of public self-supervised pathology foundation models. *Nature Communications*, 16(1):3640, 2025.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

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.