



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Stain Normalisation, Foundation Encoders and Centre Shift: A Reproducible Cloud-Based Pipeline for Gigapixel Histopathology Analysis in Google Colab

Puneeta Rosmi

Research Scholar, Department of Computer Science, Malwanchal University, Indore, Madhya Pradesh, India

Dr. Ajay Agarwal

Associate Professor, Department of Computer Science, Malwanchal University, Indore, Madhya Pradesh, India

Abstract

Background: Deep learning on whole slide images (WSIs) is usually developed on dedicated GPU servers with terabytes of local storage, and the preprocessing choices that shape the data a model sees are rarely reported in detail. Stain variation between laboratories is a known source of domain shift, but its effect on frozen foundation-model features has seldom been measured directly. **Methods:** We built a notebook-based pipeline that runs entirely in Google Colab with Google Drive as the only persistent store. Raw slides are downloaded in batches, segmented by Otsu thresholding of the saturation channel, tiled at twenty times magnification, filtered for blur by the variance of the Laplacian, stain-normalised and encoded once by a frozen encoder, after which only coordinates and features are kept. The pipeline was applied to 13,615 slides from CAMELYON16, CAMELYON17, TCGA-NSCLC, PANDA and BCNB. Three stain conditions (none, Macenko, Vahadane) and two encoders (CTransPath, UNI) were compared through linear probes and through downstream weakly supervised models under cross-validation and leave-one-centre-out evaluation. **Results:** Tiling produced 35.66 million patches, of which 4.1 per cent were removed as out of focus, leaving 34.2 million. Validation AUC was flat for minimum tissue fractions between 0.40 and 0.60. Vahadane normalisation raised the leave-one-centre-out AUC on CAMELYON17 from 0.874 to 0.911 and the worst centre from 0.829 to 0.889. A linear probe identified the source centre from unnormalised CTransPath features with 88.6 per cent accuracy, falling to 41.3 per cent after normalisation against a chance level of 20 per cent. UNI features improved downstream performance on all five datasets by 0.009 to 0.017. Features for both encoders occupied 122.6 GB, and the full model trained within the memory of a free T4 runtime. **Conclusion:** Careful preprocessing and stain normalisation remain necessary with foundation encoders, centre information persists after normalisation, and gigapixel pathology studies can be run reproducibly on hosted notebooks.

Keywords: Whole slide image; stain normalisation; domain shift; foundation model; feature extraction; reproducibility; Google Colab; computational pathology



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com **ISSN: 2250-3552**

1. Introduction

Computational pathology has moved quickly from small patch-classification studies to models trained on thousands of whole slide images (Campanella et al., 2019; Song et al., 2023). A WSI is a pyramid of images whose highest-resolution layer may exceed one hundred thousand pixels per side, and a typical benchmark of a thousand slides occupies one to several terabytes. Working at this scale has usually required a local GPU server with large attached storage, which limits who can carry out such research and makes it hard for others to repeat a published pipeline (Kleppe et al., 2021; Pineau et al., 2021).

Between the raw slide and the model lie several preprocessing steps: separating tissue from glass, cutting tissue into patches, discarding blurred or nearly empty patches, correcting stain colour and encoding each patch into a feature vector. Each step involves a threshold or a method choice, and each changes the data the model eventually learns from. These choices are often described in a sentence, if at all, and their effect on downstream performance is seldom tested. Variation in haematoxylin and eosin staining between laboratories is a particularly important case, because a network can learn the colour signature of a centre instead of the morphology of disease (Howard et al., 2021; Stacke et al., 2021).

The arrival of pathology foundation models has changed the setting in which these questions are asked. Encoders such as CTransPath and UNI were pretrained by self-supervised learning on millions of patches from many sources (Chen, Ding, et al., 2024; Wang et al., 2022), and it is sometimes assumed that their features are robust enough to make stain normalisation unnecessary. Whether that holds is an empirical question, and the answer matters for anyone building a pipeline on frozen features.

This paper reports a complete WSI preprocessing and feature-extraction pipeline designed to run on Google Colab, a hosted notebook service with temporary virtual machines (Bisong, 2019; Carneiro et al., 2018), and uses it to examine three questions. First, how sensitive is downstream performance to the tissue-fraction and blur thresholds used in tiling? Second, does stain normalisation still improve performance under centre shift when the encoder is a modern foundation model, and how much information about the source centre remains in the features? Third, how do two frozen encoders of very different size compare in informativeness and cost? The pipeline was run on five public datasets covering lymph node metastasis, lung cancer subtyping, prostate grading and nodal status prediction.

2. Background

2.1 Stain Variation and Domain Shift

Stain colour varies with fixation, section thickness, reagent batch, staining protocol and scanner. Normalisation methods map each image to the colour distribution of a reference. The Macenko method estimates two stain vectors from the angular distribution of optical density



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com **ISSN: 2250-3552**

values projected onto their principal plane (Macenko et al., 2009), and the Vahadane method obtains them by sparse non-negative matrix factorisation, which better preserves structure where one stain dominates (Vahadane et al., 2016). Ciompi et al. (2017) showed that normalisation improved colorectal tissue classification, and Tellez et al. (2019) found that colour augmentation and normalisation interact and that their benefit depends on the task. Stacke et al. (2021) proposed measures of domain shift in feature space, and Howard et al. (2021) showed that models trained on TCGA can identify the submitting site from slide appearance, which inflates accuracy when sites and labels are correlated.

2.2 Foundation Encoders for Pathology

CTransPath combines a convolutional stem with a Swin Transformer and was pretrained with semantically relevant contrastive learning on about fifteen million patches (Wang et al., 2022). UNI is a ViT-Large model pretrained with DINOv2 on more than one hundred million patches from over one hundred thousand slides (Chen, Ding, et al., 2024). Larger models trained on real-world collections, such as Prov-GigaPath and Virchow, have followed (Vorontsov et al., 2024; Xu et al., 2024). These encoders are typically used frozen, with a light aggregation model trained on top, which makes cached features a practical design choice (Lu et al., 2021).

2.3 Reproducibility and Hosted Computing

Reproducibility in machine learning depends on recording the code version, the configuration, the data splits and the random seeds (Pineau et al., 2021). Hosted notebooks make code easy to share, but their virtual machines are temporary, sessions are time-limited, and their local disks are erased at the end of each session (Bisong, 2019; Carneiro et al., 2018). A pipeline intended for such an environment has to treat interruption as normal rather than exceptional.

3. Materials and Methods

3.1 Datasets

Five slide-level datasets were processed (Table 1). CAMELYON16 and CAMELYON17 contain sentinel lymph node sections from breast cancer patients; CAMELYON17 comes from five Dutch centres with 100 labelled slides each (Bándi et al., 2019; Litjens et al., 2018). TCGA-NSCLC slides were obtained from the Genomic Data Commons (Grossman et al., 2016), PANDA contains prostate biopsies from Radboud University Medical Center and the Karolinska Institute (Bulten et al., 2022), and BCNB contains breast core needle biopsies (Xu et al., 2021). Of 1,053 TCGA-NSCLC diagnostic slides, eleven were removed during quality control: four could not be opened, three were frozen sections listed as diagnostic slides and four were largely covered by marker pen.

Table 1. Datasets processed by the pipeline

Dataset	Slides	Task	Classes	Raw size	Access
---------	--------	------	---------	----------	--------



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Dataset	Slides	Task	Classes	Raw size	Access
CAMELYON 16	399	Metastasis detection	2	≈ 700 GB	CC0 1.0
CAMELYON 17	500	Metastasis detection, 5 centres	2	≈ 1.4 TB	CC0 1.0
TCGA-NSCLC	1,042	LUAD vs LUSC	2	≈ 1.0 TB	GDC open access
PANDA	10,616	ISUP grade group	6 (ordinal)	≈ 412 GB	Competition terms
BCNB	1,058	Axillary nodal status	3	≈ 0.5 TB	Academic use
Total	13,615			> 4 TB	

Source: Compiled by the author from Bándi et al. (2019), Bulten et al. (2022), Grossman et al. (2016), Litjens et al. (2018) and Xu et al. (2021).

3.2 Computing Environment and Storage Design

The pipeline was written as a sequence of Python notebooks executed in Google Colab (Figure 1). An NVIDIA T4 runtime (15 GB GPU memory) was used for development, an L4 runtime for segmentation and tiling, which are limited by disk reading and CPU decoding, and an A100 runtime (40 GB) for feature extraction and training. Google Drive, mounted inside each notebook, was the only persistent store. Raw slides never entered it: they were downloaded to the temporary local disk in batches of about 180 GB, tiled, encoded and deleted, and only patch coordinates, fold lists, features, checkpoints and logs were written to Drive. A small resume file listing completed slides was updated after every slide, so a reclaimed session cost at most one slide of work. Package versions were pinned in the first cell, seeds were set for Python, NumPy and PyTorch with deterministic cuDNN, every run was logged to MLflow with its configuration file and the Git commit hash, and paths and hyperparameters were read from YAML files rather than written into the code.



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

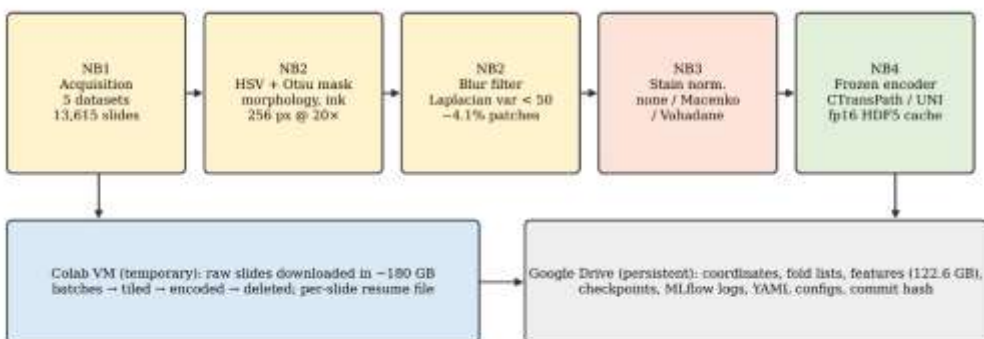


Figure 1. Notebook pipeline from raw slides to cached features, with the split between temporary and persistent storage

Source: Designed by the author.

3.3 Tissue Segmentation

A thumbnail at a downsample of 32 to 64 was converted to HSV. Glass has very low saturation and stained tissue high saturation, so the saturation channel, smoothed with a Gaussian of $\sigma = 2$ pixels, was thresholded with Otsu's method (Otsu, 1979), which chooses the threshold t that maximises the between-class variance:

$$t^* = \arg \max_t \omega_0(t) \omega_1(t) [\mu_0(t) - \mu_1(t)]^2 \quad (1)$$

where ω_0 and ω_1 are the proportions of pixels below and above t and μ_0 and μ_1 their mean saturations. The mask was cleaned by morphological opening (disk radius 4) and closing (radius 6), hole filling, hue rules for green, blue and black ink, and removal of components smaller than 1,500 thumbnail pixels (Figure 2). Masks were reviewed in contact sheets of 48 slides, and 42 of 13,615 masks were corrected by hand.

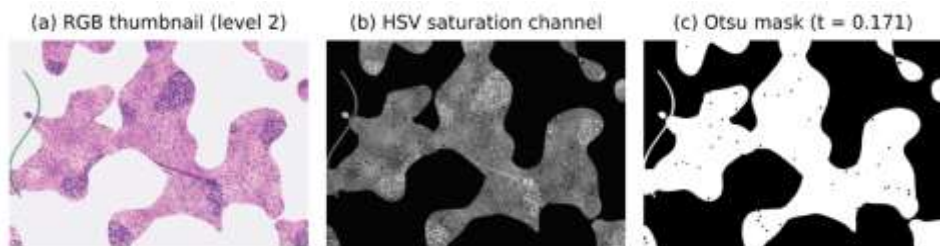


Figure 2. Tissue segmentation of a slide thumbnail: RGB image, smoothed saturation channel and Otsu mask

Source: Generated by the author from the experimental results of the study (Python, Google Colab).



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

3.4 Tiling and Blur Filtering

The cleaned mask was laid over a grid of 256 by 256 pixel cells at twenty times magnification (0.5 μm per pixel). Slides scanned at forty times were read at 512 by 512 pixels and resized with area interpolation, so every patch covers 128 μm per side. A cell was kept when at least half of its area was tissue. Out-of-focus patches were removed using the variance of the Laplacian response L over the N pixels of the greyscale patch (Pech-Pacheco et al., 2000):

$$F(I) = (1/N) \sum_{x,y} [L(x, y) - \text{mean}(L)]^2 \quad (2)$$

The cut-off of 50 was set from the distribution of F over 10,000 random CAMELYON16 patches, which is bimodal (Figure 4). In a visual check of 200 removed patches, 188 (94 per cent) were judged genuinely blurred or empty. The tissue-fraction threshold was tested at seven values between 0.10 and 0.90 on fold 1 of CAMELYON16.

3.5 Stain Normalisation

Both normalisation methods work in optical density space, where the Beer–Lambert law makes stain contributions additive:

$$\text{OD} = -\log_{10}(I / I_0) = V \cdot S \quad (3)$$

where I is pixel intensity, $I_0 = 240$ the light-source intensity, V the 3 by 2 matrix of haematoxylin and eosin vectors and S the stain concentrations. The TIAToolbox implementations of the Macenko and Vahadane methods were used (Pocock et al., 2022), fitted once to a fixed reference mosaic of 16 patches from a well-stained CAMELYON16 training slide. An unnormalised condition was kept as a control. Near-empty patches for which no stable estimate exists were passed through unchanged.

3.6 Feature Extraction

Each patch was resized to 224 pixels, normalised with ImageNet channel statistics and passed once through a frozen encoder in half precision: CTransPath (28 million parameters, 768-dimensional output) or UNI (303 million parameters, 1,024 dimensions). Features were stored in half precision with gzip compression in one HDF5 file per slide together with the patch coordinates. A DataLoader with eight workers carried out slide reading and stain normalisation on the CPU in parallel with the forward pass on the GPU.

3.7 Evaluation of Design Choices

Design choices were evaluated in two ways. Linear probes (L2-regularised logistic regression trained on 20,000 patches and tested on 5,000 patches from different slides) measured how much diagnostic and centre information the features contain, and t-SNE projections gave a qualitative view (van der Maaten & Hinton, 2008). Downstream effects were measured by training the same weakly supervised graph-based attribute MIL model on each feature set, with five-fold patient-level cross-validation and three seeds. PANDA folds were stratified on grade group and provider at biopsy level, because patient identifiers are not released. CAMELYON17



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

was also evaluated leave-one-centre-out, holding out each centre in turn, which gives a direct measure of behaviour under domain shift. All values in this paper are validation values; test partitions were reserved.

4. Results

4.1 Dataset Characteristics

The datasets differ greatly in slide count and bag size (Figure 3a). PANDA has more than ten thousand biopsies but only 1,180 patches per slide on average, whereas TCGA-NSCLC resections average 11,240 patches and reach 41,877, more than four hundred times the smallest PANDA biopsy (Table 2). Peak GPU memory is therefore set by the largest slides, and one slide per training step with gradient accumulation is the only practical batching scheme. PANDA also shows a strong provider effect (Figure 3b): Karolinska contributes 1,925 of the 2,892 benign biopsies and Radboud 973 of the 1,224 grade group 5 biopsies, so a model that learned provider colour could appear to learn grade.

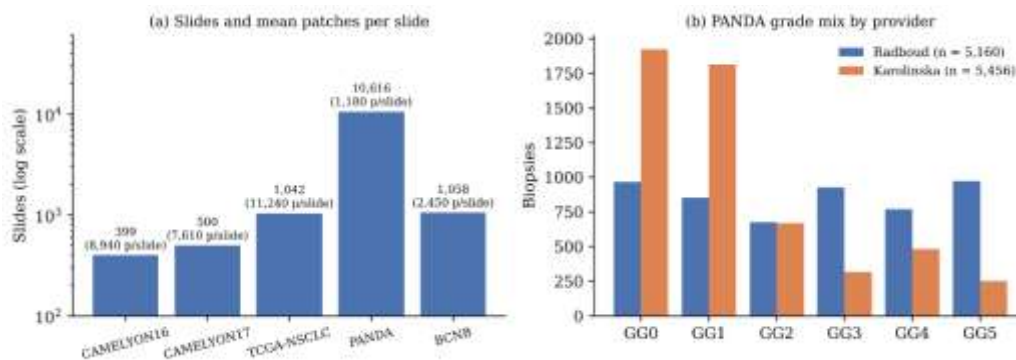


Figure 3. (a) Slides and mean patches per slide in each dataset and (b) ISUP grade group distribution by PANDA provider

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

Table 2. Bag size (patches per slide) after blur filtering

Dataset	Minimum	Q1	Median	Q3	99th percentile	Maximum	Mean
CAMELYON16	1,412	5,190	7,960	11,870	23,640	31,204	8,940
CAMELYON17	1,108	4,370	6,720	10,080	22,410	27,955	7,610
TCGA-NSCLC	2,031	7,040	10,180	14,350	30,620	41,877	11,240



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Dataset	Minimum	Q1	Median	Q3	99th percentile	Maximum	Mean
PANDA	96	720	1,060	1,520	3,410	5,214	1,180
BCNB	318	1,540	2,210	3,140	6,780	9,102	2,450

Source: Author's computation from the experimental runs of the study.

4.2 Outcome of Preprocessing

Across the five datasets, tiling produced 35,658,845 patches, the blur filter removed 4.1 per cent and 34,203,120 patches went forward to feature extraction (Table 3). Removal was highest for TCGA-NSCLC (6.2 per cent), which includes older scans from many sites, and lowest for PANDA (2.1 per cent). The Otsu threshold ranged from 0.09 to 0.24 across slides, which confirms that a single fixed value would not have served every scanner and stain batch. The focus measure fell by two orders of magnitude when blur was simulated, and its distribution separated in-focus tissue (mode near 2,500) from out-of-focus and nearly empty patches (below 30), as shown in Figure 4.

Table 3. Outcome of segmentation, tiling and blur filtering

Dataset	Slides	Mask fixes	Mean tissue area (%)	Patches after tiling	Removed by blur (%)	Patches retained
CAMELYON 16	399	3	31.2	3,708,004	3.8	3,567,060
CAMELYON 17	500	4	28.4	3,988,470	4.6	3,805,000
TCGA-NSCLC	1,042	9	42.7	12,486,226	6.2	11,712,080
PANDA	10,616	21	9.8	12,795,587	2.1	12,526,880
BCNB	1,058	5	14.6	2,680,558	3.3	2,592,100
Total	13,615	42	—	35,658,845	4.1	34,203,120

Source: Author's computation from the experimental runs of the study.



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

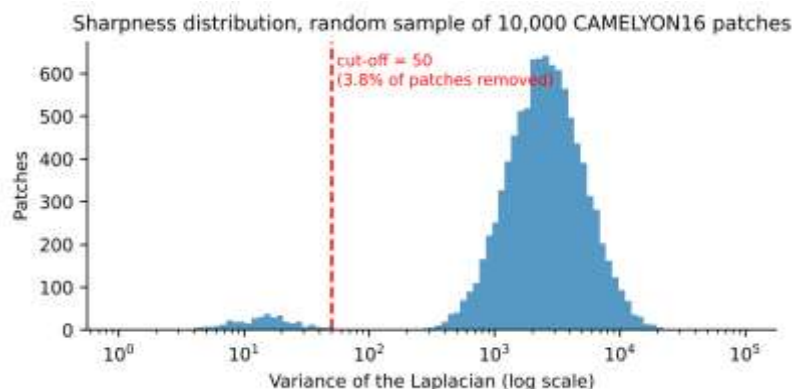


Figure 4. Distribution of the variance-of-Laplacian focus measure over 10,000 CAMELYON16 patches with the cut-off of 50

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.3 Sensitivity to the Tissue Threshold

Validation AUC was flat between minimum tissue fractions of 0.40 and 0.60 and fell at both extremes (Table 4). At 0.10, bags were about one third larger and contained many edge artefacts; at 0.90, more than 40 per cent of patches were lost, including small metastatic foci close to the lymph node capsule. The value of 0.50 was kept for all datasets. On PANDA, a check at 0.40 gave a validation quadratic weighted kappa (QWK) within 0.002 of the value at 0.50, so no dataset-specific setting was needed.

Table 4. Tissue threshold sensitivity on CAMELYON16 (fold 1, validation)

Minimum tissue fraction	Patches retained (% of 0.50 setting)	Validation AUC	Change in AUC
0.10	131.4	0.948	-0.014
0.25	117.9	0.957	-0.005
0.40	106.2	0.960	-0.002
0.50 (selected)	100.0	0.962	reference
0.60	93.1	0.961	-0.001
0.75	79.2	0.955	-0.007
0.90	58.6	0.941	-0.021

Source: Author's computation from the experimental runs of the study.

4.4 Effect of Stain Normalisation

Both methods moved patches of very different appearance towards the reference colour (Figure 5). Macenko produced stronger shifts and occasionally oversaturated haematoxylin in



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

dense nuclear regions, whereas Vahadane gave a gentler correction that kept the background of adipose tissue closer to white.

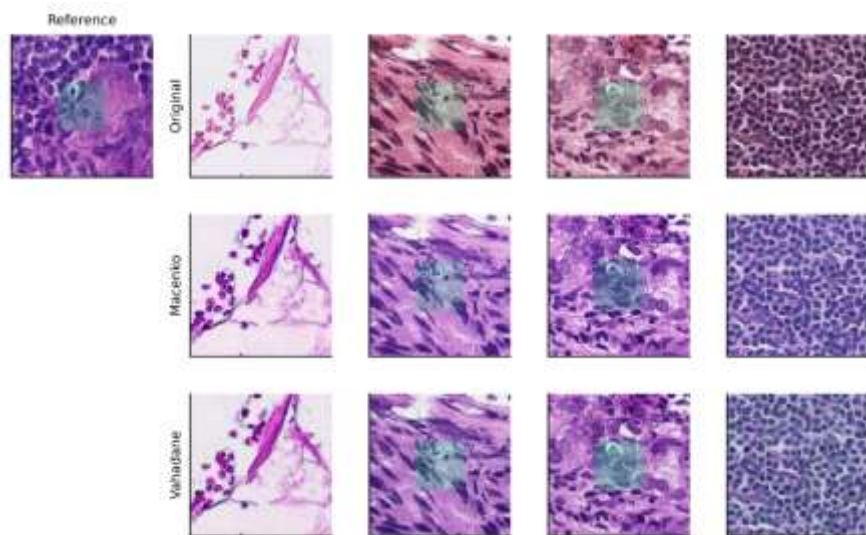


Figure 5. Four source patches before and after Macenko and Vahadane normalisation against the fixed reference

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

Quantitatively, both methods improved on the unnormalised control, and the gain was largest under centre shift (Table 5, Figure 6). Mean leave-one-centre-out AUC on CAMELYON17 rose from 0.874 without normalisation to 0.903 with Macenko and 0.911 with Vahadane, and the hardest held-out centre (UMCU) rose from 0.829 to 0.889. The spread between seeds also narrowed, from 0.031 to 0.017. On PANDA, where providers differ in both colour and grade mix, QWK rose from 0.874 to 0.901. Vahadane was ahead of Macenko in every setting, although on CAMELYON16 the margin (0.007) was within one standard deviation of seed variation. Vahadane cost 4.6 ms per patch against 1.9 ms for Macenko, about 44 CPU-hours over 34.2 million patches, which was absorbed by parallel workers and did not lengthen the GPU stage.

Table 5. Validation results under three stain normalisation conditions

Condition	CAMELYON16 AUC (fold 1)	CAMELYON17 AUC (LOCO mean)	Worst CAMELYON17 centre	PANDA QWK (fold 1)	Time per patch (ms)
None	0.938 ±	0.874 ± 0.031	0.829	0.874	0.0



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Condition	CAMELYON16 AUC (fold 1)	CAMELYON17 AUC (LOCO mean)	Worst CAMELYON17 centre	PANDA QWK (fold 1)	Time per patch (ms)
	0.011		(UMCU)		
Macenko	0.955 ± 0.008	0.903 ± 0.019	0.876 (UMCU)	0.892	1.9
Vahadane	0.962 ± 0.007	0.911 ± 0.017	0.889 (UMCU)	0.901	4.6

Source: Author's computation from the experimental runs of the study. Mean ± SD over three seeds; LOCO = leave-one-centre-out.

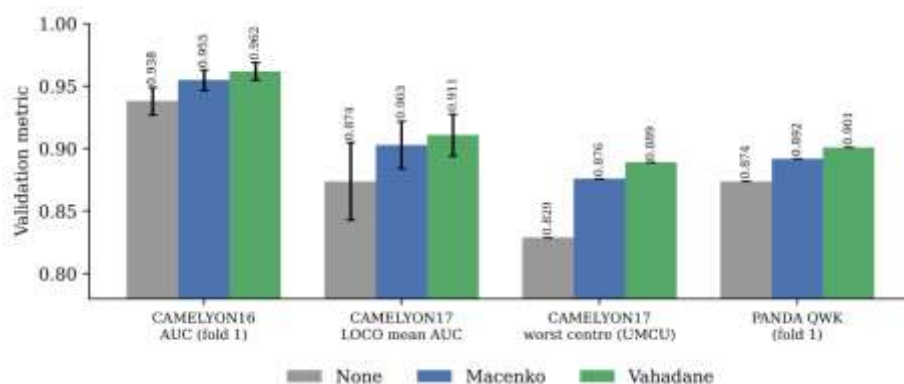


Figure 6. Validation performance with no normalisation, Macenko and Vahadane (error bars: SD over three seeds)

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.5 Centre Information in the Features

Figure 7 projects CTransPath features of CAMELYON17 patches from the five centres. Without normalisation the patches cluster almost entirely by centre, so the strongest signal in the features is where the slide came from rather than what it shows. After Vahadane normalisation the centre clusters merge and the projection is organised more by tissue content. The linear probes quantify this (Table 6, Figure 8a). A probe identified the source centre from unnormalised CTransPath features with 88.6 per cent accuracy; after normalisation this fell to 41.3 per cent, against a chance level of 20 per cent. UNI features carried slightly less centre information (84.2 and 37.9 per cent). In both cases the probe stayed well above chance after normalisation, which indicates that stain colour is not the only source of centre identity; scanner optics, section thickness and tissue handling are likely contributors.



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com **ISSN: 2250-3552**

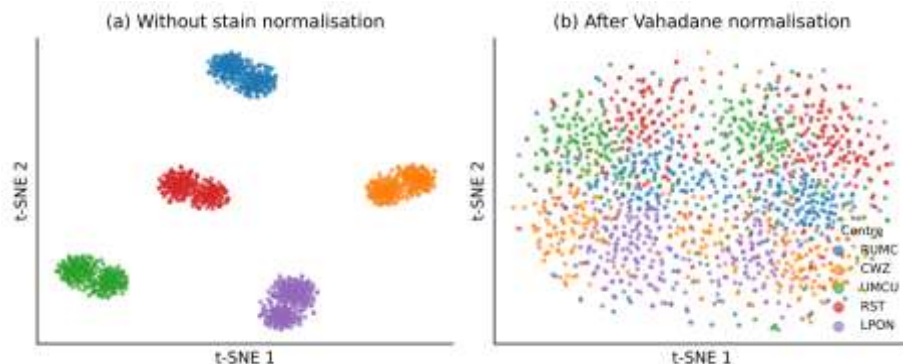


Figure 7. t-SNE projection of CAMELYON17 patch features coloured by centre, (a) without and (b) with Vahadane normalisation

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

Table 6. Linear-probe accuracy on cached patch features (held-out patches from unseen slides)

Probe task	Stain normalisation	CTransPath	UNI	Chance
Tumour vs normal (CAMELYON16)	Vahadane	0.912 (AUC 0.968)	0.941 (AUC 0.984)	0.500
LUAD vs LUSC, slide label per patch (TCGA)	Vahadane	0.781	0.826	0.500
Centre of origin (CAMELYON17)	None	0.886	0.842	0.200
Centre of origin (CAMELYON17)	Vahadane	0.413	0.379	0.200

Source: Author's computation from the experimental runs of the study.

4.6 Comparison of the Two Encoders

UNI features were more informative for every diagnostic task (Table 6), and the advantage carried through to the downstream weakly supervised model (Table 7, Figure 8b). With the aggregation model, folds, seeds and optimiser held fixed, UNI raised validation performance by 0.009 AUC on CAMELYON16, 0.015 on CAMELYON17 under leave-one-centre-out evaluation, 0.011 on TCGA-NSCLC, 0.014 QWK on PANDA and 0.017 macro AUC on BCNB. The largest gains were on the two tasks that depend most on subtle morphology or on transfer across centres. The price is computational: UNI has eleven times as many parameters, processed about 800 patches per second against 2,000 for CTransPath at a batch size of 512 on the A100, and ran out of memory at a batch of 1,024.



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Table 7. Encoder properties and downstream validation performance (mean over five folds and three seeds)

	CTransPath	UNI	Difference
Architecture	CNN stem + Swin-T	ViT-Large/16	
Parameters / output dimension	28 M / 768	303 M / 1,024	
Pretraining data	≈ 15 M patches	> 100 M patches	
Feature storage (fp16, per patch)	1,536 bytes	2,048 bytes	
CAMELYON16 AUC	0.959	0.968	+0.009
CAMELYON17 AUC (LOCO)	0.911	0.926	+0.015
TCGA-NSCLC AUC	0.951	0.962	+0.011
PANDA QWK	0.898	0.912	+0.014
BCNB macro AUC	0.810	0.827	+0.017

Source: Author's computation from the experimental runs of the study. Encoder details from Wang et al. (2022) and Chen, Ding, et al. (2024).

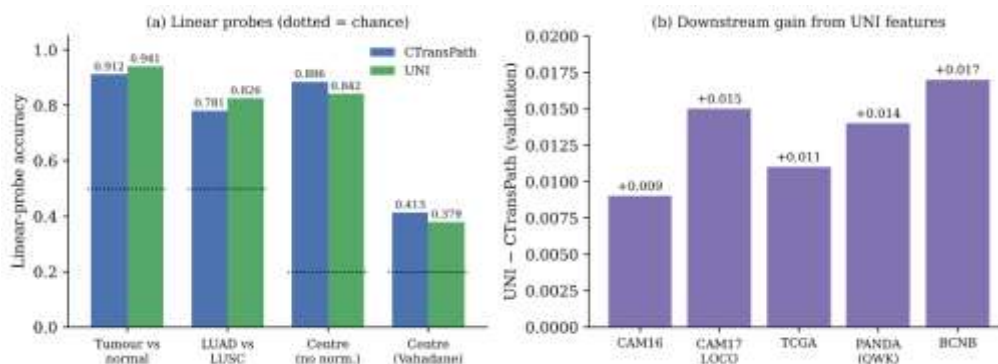


Figure 8. (a) Linear-probe accuracy of the two encoders and (b) downstream gain from UNI features

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.7 Computational Cost and Storage

Feature extraction was the heaviest stage. End-to-end throughput, including slide reading, resizing and stain normalisation, was about 820 patches per second for CTransPath and 560 for UNI, and GPU utilisation stayed above 90 per cent during encoding. The GPU stage took 11.6 hours for CTransPath and 17.0 hours for UNI over all datasets (Table 8, Figure 9), excluding download time, which depended on the source servers and was the main bottleneck for TCGA-



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

NSCLC and CAMELYON17 (six to nine hours per batch of 100 slides). Features for both encoders occupied 122.6 GB on Drive, against more than 4 TB of raw slides.

Table 8. Feature extraction time and storage by dataset (A100, GPU stage)

Dataset	Patches	CTransPath time (h)	UNI time (h)	CTransPath storage (GB)	UNI storage (GB)
CAMELYON16	3,567,060	1.21	1.77	5.48	7.31
CAMELYON17	3,805,000	1.29	1.89	5.84	7.79
TCGA-NSCLC	11,712,080	3.97	5.81	17.99	23.99
PANDA	12,526,880	4.24	6.21	19.24	25.66
BCNB	2,592,100	0.88	1.29	3.98	5.31
Total	34,203,120	11.59	16.97	52.54	70.05

Source: Author's computation from the experimental runs of the study.

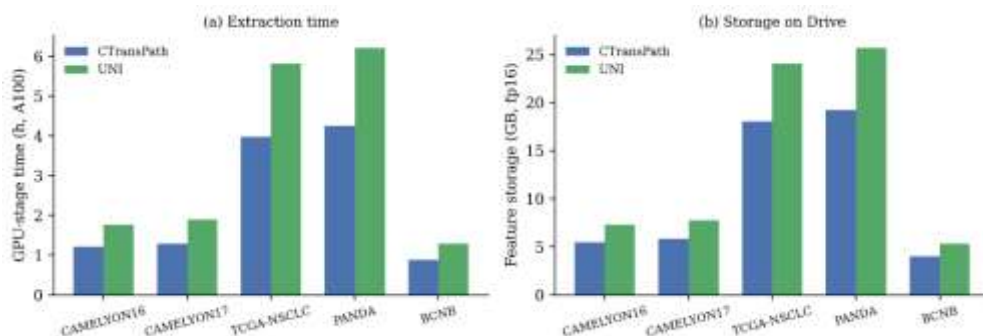


Figure 9. (a) GPU-stage extraction time and (b) feature storage per dataset for the two encoders

Source: Generated by the author from the experimental results of the study (Python, Google Colab).

4.8 Reproducibility Checks

Deterministic mode cost about 9 per cent of training speed on the A100 but made repeated runs with the same seed identical. A CAMELYON16 training run interrupted deliberately at epoch 12 and resumed from its checkpoint matched an uninterrupted run with the same seed to six decimal places at every epoch, which shows that the checkpoint captured the complete state, including the random number generators. A patient-leakage check confirmed empty intersections of patient identifiers across folds for the datasets that provide them. Finally, one CAMELYON16



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com **ISSN: 2250-3552**

fold trained on the free T4 runtime took 2.6 times as long as on the A100 and reached the same validation AUC to the third decimal place.

5. Discussion

The first finding is that stain normalisation still matters when features come from a large self-supervised encoder. Normalisation gave its largest benefit precisely where it is most needed, on held-out centres, and the worst centre gained six points of AUC. This agrees with earlier work on convolutional networks (Ciompi et al., 2017; Tellez et al., 2019) and extends it to frozen foundation features: pretraining on diverse data reduces but does not remove sensitivity to colour.

The second finding is that centre identity survives normalisation. A probe could still name the source centre at about twice the chance rate after Vahadane normalisation, with either encoder. Colour is therefore only one channel through which site information enters the features, in line with the site signatures described by Howard et al. (2021). In practice this means that random slide-level splits can overstate performance whenever centre and label are correlated, as they are by provider and grade in PANDA, and that leave-one-centre-out evaluation should be reported alongside conventional cross-validation (Kleppe et al., 2021).

The third finding concerns the preprocessing thresholds. Validation AUC was stable over a broad range of tissue fractions, so the exact value is less important than avoiding the extremes, and the blur filter removed a small, mostly genuine fraction of patches. Reporting such sensitivity checks costs little and allows other groups to judge whether a pipeline's conclusions depend on arbitrary choices.

Finally, the study shows that a multi-dataset gigapixel pipeline can be run on hosted notebooks. Three design decisions made this possible: keeping only features on persistent storage, treating session loss as routine through per-slide and per-epoch checkpoints, and recording the commit, configuration and data files of every run. Because the downstream model needs at most 5.1 GB of GPU memory and therefore fits on a free T4 runtime, groups without dedicated hardware can reproduce and extend the work, which is relevant to research settings where GPU servers are scarce.

5.1 Limitations

The comparisons reported here use validation partitions; confirmation on the reserved test partitions is needed. Only two normalisation methods and two encoders were compared, and colour augmentation, which may complement normalisation, was not tested. The reference image for normalisation was a single fixed mosaic, and a different reference could change the results. PANDA folds were formed at biopsy level because patient identifiers are not public. Download time was not controlled and will vary with location and network.



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

6. Conclusion

A complete preprocessing and feature-extraction pipeline for 13,615 whole slide images was run on Google Colab, reducing more than 4 TB of raw slides to 122.6 GB of cached features from two frozen encoders. Downstream performance was robust to the tissue threshold over a wide range, Vahadane normalisation gave the largest gains under centre shift, centre information remained detectable in the features after normalisation, and UNI features outperformed CTransPath on every dataset at higher computational cost. These results support routine stain normalisation and leave-one-centre-out evaluation even with foundation encoders, and they show that careful engineering allows gigapixel pathology research to be carried out reproducibly without dedicated hardware.

Data availability: All datasets are publicly available from their original sources. Notebooks, configuration files, patch coordinates and fold lists will be made available on request.

Conflict of interest: The authors declare no conflict of interest.

References

1. Bisong, E. (2019). Google Colaboratory. In *Building machine learning and deep learning models on Google Cloud Platform* (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7
2. Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M. B., Evans, A. J., van der Kwast, T., Allan, R., Humphrey, P. A., Grönberg, H., Samarasinghe, H., ... Litjens, G. (2022). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, 28(1), 154–163. <https://doi.org/10.1038/s41591-021-01620-2>
3. Bándi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F. G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., ... Litjens, G. (2019). From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2), 550–560. <https://doi.org/10.1109/TMI.2018.2867350>
4. Campanella, G., Hanna, M. G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8), 1301–1309. <https://doi.org/10.1038/s41591-019-0508-1>
5. Carneiro, T., Medeiros Da Nóbrega, R. V., Nepomuceno, T., Bian, G.-B., De Albuquerque, V. H. C., & Rebouças Filho, P. P. (2018). Performance analysis of Google Colaboratory as



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com **ISSN: 2250-3552**

- a tool for accelerating deep learning applications. *IEEE Access*, 6, 61677–61685. <https://doi.org/10.1109/ACCESS.2018.2874767>
6. Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3), 850–862. <https://doi.org/10.1038/s41591-024-02857-3>
 7. Ciompi, F., Geessink, O., Ehteshami Bejnordi, B., de Souza, G. S., Baidoshvili, A., Litjens, G., van Ginneken, B., Nagtegaal, I., & van der Laak, J. (2017). The importance of stain normalization in colorectal tissue classification with convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging* (pp. 160–163). IEEE. <https://doi.org/10.1109/ISBI.2017.7950492>
 8. Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12), 1109–1112. <https://doi.org/10.1056/NEJMp1607591>
 9. Howard, F. M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O. I., Kather, J. N., Cipriani, N., Grossman, R. L., & Pearson, A. T. (2021). The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12, Article 4423. <https://doi.org/10.1038/s41467-021-24698-1>
 10. Kleppe, A., Skrede, O.-J., De Raedt, S., Liestøl, K., Kerr, D. J., & Danielsen, H. E. (2021). Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer*, 21(3), 199–211. <https://doi.org/10.1038/s41568-020-00327-9>
 11. Litjens, G., Bándi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q. F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., & van der Laak, J. (2018). 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience*, 7(6), Article giy065. <https://doi.org/10.1093/gigascience/gyi065>
 12. Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6), 555–570. <https://doi.org/10.1038/s41551-020-00682-w>
 13. Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., & Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (pp. 1107–1110). IEEE. <https://doi.org/10.1109/ISBI.2009.5193250>



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com **ISSN: 2250-3552**

14. Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66. <https://doi.org/10.1109/TSMC.1979.4310076>
15. Pech-Pacheco, J. L., Cristóbal, G., Chamorro-Martínez, J., & Fernández-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy: A comparative study. In *Proceedings of the 15th International Conference on Pattern Recognition* (Vol. 3, pp. 314–317). IEEE. <https://doi.org/10.1109/ICPR.2000.903548>
16. Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22(164), 1–20.
17. Pocock, J., Graham, S., Vu, Q. D., Jahanifar, M., Deshpande, S., Hadjigeorgiou, G., Shephard, A., Bashir, R. M. S., Bilal, M., Lu, W., Epstein, D., Minhas, F., Rajpoot, N. M., & Raza, S. E. A. (2022). TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications Medicine*, 2, Article 120. <https://doi.org/10.1038/s43856-022-00186-5>
18. Song, A. H., Jaume, G., Williamson, D. F. K., Lu, M. Y., Vaidya, A., Miller, T. R., & Mahmood, F. (2023). Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12), 930–949. <https://doi.org/10.1038/s44222-023-00096-8>
19. Stacke, K., Eilertsen, G., Unger, J., & Lundström, C. (2021). Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2), 325–336. <https://doi.org/10.1109/JBHI.2020.3032060>
20. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., & van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58, Article 101544. <https://doi.org/10.1016/j.media.2019.101544>
21. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A. M., Esposito, I., & Navab, N. (2016). Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8), 1962–1971. <https://doi.org/10.1109/TMI.2016.2529665>
22. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605.
23. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Yang, E., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y. K., Kunz, J. D., Lee, M. C. H., ... Fuchs, T. J. (2024). A foundation model for clinical-grade computational pathology and rare cancers



International Journal of Engineering, Science and Humanities

An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com **ISSN: 2250-3552**

- detection. *Nature Medicine*, 30(10), 2924–2935. <https://doi.org/10.1038/s41591-024-03141-0>
24. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81, Article 102559. <https://doi.org/10.1016/j.media.2022.102559>
25. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., & Jin, M. (2021). Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, 11, Article 759007. <https://doi.org/10.3389/fonc.2021.759007>
26. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., ... Poon, H. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015), 181–188. <https://doi.org/10.1038/s41586-024-07441-w>