

Evaluating Vision Transformers for Stain Robustness and Calibration in Low-Resource Digital Pathology

Abir Hasan

International University of Business Agriculture and Technology

Sadia Yasmin

International University of Business Agriculture and Technology

Ripa Akter

International University of Business Agriculture and Technology

Md. Abdul Awal

`abdul.awal@iubat.edu`

International University of Business Agriculture and Technology

Research Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11111456/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Evaluating Vision Transformers for Stain Robustness and Calibration in Low-Resource Digital Pathology

Abir Hasan¹, Sadia Yasmin¹, Ripa Akter¹, Md. Abdul Awal^{1*}

^{1*}Department of Computer Science and Engineering, International
University of Business Agriculture and Technology (IUBAT), Dhaka,
Bangladesh.

*Corresponding author(s). E-mail(s): abdul.awal@iubat.edu;

Abstract

Deep learning models for histopathology can achieve high in-domain discrimination while remaining vulnerable to color and staining variation and probability miscalibration, limiting reliable use under distribution shift. This paper evaluates, under a fixed data budget and common ImageNet pretraining, how CNN and Vision Transformer backbones differ in discrimination, RGB color-shift robustness, and probability calibration. Under a constrained, reproducible regime – 20,000 PatchCamelyon (PCam) patches (8% of the benchmark), five random seeds, a single-GPU Google Colab environment – we compare five ImageNet-pretrained backbones (ResNet18, ResNet50, EfficientNet-B0, DenseNet121, and a plain ViT-B/16) with *no* pathology-specific pretraining, so pretraining-data scale cannot explain any observed difference; the backbones differ in parameter count and learning rate, so the comparison reflects an architecture-scale association, not an isolated causal effect. All five backbones score within a few AUC points in-domain (ViT-B/16 highest, 0.9634 ± 0.0076), but at $\sigma = \mathbf{0.20}$ ViT-B/16 shows only 1.5% relative AUC degradation under RGB channel-gain corruption versus 7.7–11.0% for the CNNs (Holm-corrected $p < \mathbf{0.005}$), replicated under a second, stain-specific hematoxylin-eosin perturbation, while also achieving the lowest in-domain Expected Calibration Error (0.044 vs. 0.071–0.094) and Brier score. Zero-shot evaluation on three independent external datasets – BACH and PathMNIST (cross-task transfer) and CAMELYON17-WILDS (same-task, cross-hospital transfer) – surfaces a model- and protocol-specific failure mode on the cross-task datasets, where one CNN collapses to all-positive predictions while the transformer degrades gracefully; on CAMELYON17-WILDS all five backbones generalize comparably (AUC 0.988–0.992), indicating the robustness and

calibration advantages are clearest under stain/color perturbation and cross-task transfer rather than under general domain shift. A two-level explainability pipeline, full computational-cost characterization, a gated multimodal fusion module, and an HL7 FHIR R4 export stage complete the framework, with full detail in the Supplementary Material. We report this as a reusable low-resource benchmark protocol for computationally constrained research settings.

Keywords: digital pathology, vision transformer, model calibration, stain robustness, external validation, low-resource deep learning

1 Introduction

Whole-slide imaging (WSI) has turned histopathology into a digital, computationally tractable data source, and its integration with artificial intelligence (AI) is now recognized as a central direction for anatomic pathology [1, 2]. Deep convolutional and transformer-based models report near-expert diagnostic accuracy on tasks such as breast lymph node metastasis detection [3, 4], and reviews spanning WSI acquisition, preprocessing, and downstream deep learning consistently identify the same open problems: sensitivity to pre-analytical variation (tissue preparation, staining, scanner artifacts), high computational demand, and limited generalization across sites and settings [5–7]. A large-scale systematic review and meta-analysis of AI diagnostic-accuracy studies in digital pathology further finds that most published evaluations carry substantial risk of bias and weak external validation [8], underscoring that reported accuracy figures alone are not a reliable guide to clinical readiness. The great majority of recent progress, moreover, has been demonstrated under conditions that are difficult to reproduce outside well-resourced research groups: full-scale datasets of hundreds of thousands to millions of whole-slide images, multi-GPU training pipelines, and, increasingly, foundation models pretrained on tens of millions of pathology tiles with hundreds of millions to billions of parameters [9–12]. For hospitals and research groups operating with limited compute – a common situation outside a small number of well-funded centers – this creates a widening gap between published state-of-the-art (SOTA) performance and what is practically deployable.

At the same time, raw classification accuracy is not the primary obstacle to clinical adoption of AI-assisted pathology. Two other properties matter at least as much for real-world use. (i) *Robustness* to the substantial visual variability introduced by different scanners, laboratories, and hematoxylin-and-eosin (H&E) staining protocols is well documented to degrade convolutional neural network (CNN) performance even when accuracy on a single site’s data is high [13]. (ii) *Calibration*, i.e., whether a model’s predicted probabilities reflect true likelihoods, determines whether a clinician can trust a borderline score. Multimodal, externally validated models that combine imaging with clinical variables are also rare outside large consortium studies; a recent post-hoc analysis of four phase 3 randomised trials under the STAMPEDE platform protocol is a notable exception, showing that a locked multimodal AI prognostic model combining

digitized biopsy images with clinical variables remained significantly prognostic when externally validated on independent trial cohorts [15].

This paper studies robustness and calibration directly, using a deliberately constrained experimental regime – 20,000 training patches (8% of the full PatchCamelyon training set) on a single-GPU Google Colab environment, five independent random seeds per model – to reflect the resource envelope available to most non-foundation-model research groups working under comparable compute constraints. Within this regime, we compare four CNN backbones (ResNet18, ResNet50, EfficientNet-B0, DenseNet121) against a Vision Transformer (ViT-B/16) [16] pretrained only on ImageNet, with no pathology-specific self-supervised pretraining. Our central finding is that the transformer’s advantage in this setting is not primarily raw accuracy – all five backbones score within a few AUC points of one another – but a substantially smaller degradation under stain perturbation and better-calibrated probabilities, quantified in Section 5.

We embed this finding in a deployment-oriented, low-resource research pipeline – a two-level explainability output (bag-level MIL attention plus patch-level Grad-CAM) and a full computational-cost characterization – and additionally implement two framework-demonstration modules, a gated multimodal fusion architecture and an HL7 FHIR R4 export stage, whose complete detail we place in the Supplementary Material so that the main text stays focused on the central evaluation.

We do not claim to match the raw accuracy of billion-parameter pathology foundation models such as UNI [10] or Prov-GigaPath [11], which are pretrained on orders of magnitude more pathology-specific data than we use here; we report their published numbers alongside ours in Section 6 for honest context. Instead, we argue that stain robustness and calibration – properties that determine whether a model survives contact with a real, multi-site clinical workflow – are underserved by the current focus on leaderboard accuracy, and that a Vision Transformer trained under modest compute constraints already offers a meaningfully more deployable trade-off than CNNs of comparable or larger size.

1.1 Contributions

This paper makes *one* central contribution, supported by two pieces of infrastructure that make it trustworthy and actionable rather than by a list of independent novelties.

Central contribution – under matched data and pretraining conditions, transformer architecture is associated with greater robustness and calibration under distribution shift. Holding training data, budget, and evaluation protocol fixed, ViT-B/16 degrades only 1.5% relative AUC under a $\sigma = 0.2$ custom RGB channel-gain corruption motivated by stain/color-shift robustness studies [13] versus 7.7–11.0% for four CNN backbones, while simultaneously producing the best-calibrated probabilities (Expected Calibration Error, Brier score) of the five models – despite using only ImageNet pretraining, so pathology-specific data scale cannot explain the gain. Because ViT-B/16 also differs from the CNNs in parameter count and learning rate, and only one transformer architecture is evaluated, this is an architecture-scale association demonstrated under a fixed, controlled protocol, not a

fully isolated causal effect of architecture alone (Section 7). We additionally evaluate cross-dataset, cross-task zero-shot transfer on two independent external datasets (BACH; the PathMNIST test split derived from CRC-VAL-HE-7K), with quantitative error analysis that surfaces a model- and protocol-specific degenerate failure mode in one CNN – an evaluation we report as a distinct, more heavily confounded transfer setting rather than as a direct extension of the stain-perturbation result (Section 7) – and a third, same-task, cross-hospital external dataset (CAMELYON17-WILDS) on which all five backbones generalize comparably well, which we report to scope the robustness claim honestly rather than to inflate its apparent generality.

This claim rests on two supporting components, which we present as enabling infrastructure rather than as separate contributions. *(i) A reusable low-resource benchmark protocol:* the controlled five-backbone, five-seed comparison on PatchCamelyon under a constrained budget (20,000 patches, single-GPU Google Colab environment), with a statistical stack – DeLong and paired- t tests, bootstrap confidence intervals, McNemar’s test, and effect sizes (Sections 3.10, 5) – and a data-efficiency analysis showing the budget is not accuracy-limiting, so the comparison is reproducible without institutional-scale hardware. *(ii) A deployment wrapper:* a two-level explainability pipeline (bag-level MIL attention with Grad-CAM/Grad-CAM++, evaluated by insertion/deletion faithfulness) and a full computational-cost characterization (Section 3.11) that ground the “low-resource” claim in measured numbers. Two additional framework-demonstration modules – a gated multimodal fusion architecture and an HL7 FHIR R4 export pipeline – are implemented and evaluated in full but, being demonstrations rather than the paper’s claim, are documented entirely in the Supplementary Material.

The remainder of this paper is organized as follows. Section 2 reviews related work in CNN and transformer-based pathology classification, stain robustness, and calibration. Section 3 describes the dataset, architectures, training protocol, and evaluation methodology. Section 4 details the experimental design. Section 5 reports quantitative results for backbone comparison, calibration, robustness, external validation, and explainability. Section 6 discusses clinical implications and positions our results against modern foundation models. Section 7 states limitations and future work. Section 8 concludes.

2 Related Work

2.1 CNN and Transformer Architectures for Histopathology Classification

Convolutional networks have been the default for histopathology patch and slide classification for a decade. The PatchCamelyon (PCam) benchmark used here derives from the CAMELYON16 lymph-node metastasis challenge [4] and was introduced by [3] with a rotation-equivariant CNN (P4M-DenseNet, 0.963 AUC). The Vision Transformer (ViT) [16] has since been adopted across computational pathology, both as a standalone classifier and as the backbone of foundation models pretrained at a scale distinct from task-specific classifiers: CTransPath [9] (Swin Transformer, 15M patches), UNI [10] (100M+ patches, 100K+ slides), and Prov-GigaPath [11]

(1.3B tiles, 171K slides). This scale delivers strong accuracy but is inaccessible without proprietary slide archives and multi-GPU/TPU clusters, and reviews of WSI-based AI repeatedly flag limited compute and cross-site generalization – not architecture choice – as the dominant deployment barriers [5–7, 12]. We therefore use a *plain*, ImageNet-only ViT-B/16 as one of five directly comparable backbones. Using ImageNet-pretrained models removes pathology-specific pretraining scale as an explanation for any observed robustness or calibration advantage; however, architecture, parameter count, and architecture-specific optimization remain jointly varied across the five backbones (Section 7), and the comparison is reproducible without institutional-scale infrastructure.

2.2 Robustness to Stain and Scanner Variation

Variation in H&E staining, reagent batch, and scanner hardware is a well-established source of domain shift, and [13] established the standard practice of quantifying robustness through synthetic stain-color perturbation rather than single-site accuracy. A systematic review and meta-analysis found most pathology-AI evaluations carry substantial risk of bias and weak external validation [8], and WSI reviews treat staining and scanner artifacts as a recurring failure mode [5–7]. Little of this literature isolates *architecture* as the robustness variable – most studies vary augmentation or normalization on a single backbone. We instead hold a custom RGB channel-gain corruption protocol – motivated by, but not identical to, the stain-color augmentation studies of [13] – fixed across backbones and vary the backbone. Holding the corruption procedure fixed permits a controlled comparison of model sensitivity to the same perturbation; causal attribution to architecture alone would additionally require capacity- and optimization-matched ablations (Section 7).

2.3 Calibration of Medical Deep Learning Models

Modern deep networks are frequently overconfident: [19] showed that greater depth and width improve accuracy yet often *worsen* calibration. In a medical setting a miscalibrated probability is a failure mode distinct from misclassification – a confidently wrong score is more dangerous than an uncertain correct one – yet calibration is reported far less consistently than AUC in the pathology AI literature. We report Expected Calibration Error and Brier score alongside AUC for all five backbones, treating calibration as a first-class metric.

2.4 Multiple-Instance Learning, Multimodal Fusion, Explainability, and Interoperability

Because slide-level annotations are costly, whole-slide classification is commonly formalized as multiple-instance learning (MIL), treating a slide as a “bag” of patch instances with only a bag-level label; [20] introduced learnable gated attention-pooling as a differentiable, interpretable alternative to fixed pooling, which we benchmark against mean/max pooling under our low-resource regime (Section 5). For patch-level explainability, Grad-CAM [21] and Grad-CAM++ [24] are the most widely used techniques but are typically shown as qualitative heatmaps; quantitative faithfulness

(insertion/deletion AUC) is reported far less often, and reviews note that explainability is claimed more than it is measured [6, 12]. Two further directions motivate our supplementary demonstrations: multimodal fusion of imaging with structured clinical variables, identified as promising but underused [2, 12] and shown by STAMPEDE [15] to yield externally validated clinical utility; and standards-based reporting, where HL7 FHIR [23] – supported by most hospital EHRs – is essentially absent from a pathology AI literature that usually stops at a raw probability, mirroring the ongoing DICOM transition on the imaging side [14, 22].

2.5 Positioning of This Work

No single work in the literature reviewed combines a controlled, low-resource CNN-versus-transformer comparison; a quantified, controlled cross-architecture color-robustness comparison; systematic calibration reporting; and multi-dataset external validation in one reproducible framework. A point-by-point summary of the specific focus and gap addressed for each work discussed above is given in Supplementary Material Section S7. Section 3 describes how each of these components is implemented and evaluated.

3 Methodology

3.1 Framework Overview

Figure 1 summarizes the end-to-end framework. A patch is encoded by one of five backbone architectures (Section 3.3); at synthetic bag level (PCam provides no native slide grouping, Section 3.7), patch embeddings are aggregated by a multiple-instance-learning (MIL) pooling operator (Section 3.7); the resulting representation yields a prediction together with a calibration assessment (Section 3.5) and a two-level explainability output (Section 3.9). Two evaluation-time branches – stain-perturbation robustness testing and zero-shot external-dataset validation, and CNN-layer Grad-CAM/Grad-CAM++ – reuse the trained backbone directly, without additional fine-tuning, and are shown as dashed boxes. Two further deployment components, a gated multimodal fusion module and an HL7 FHIR R4 export stage, sit downstream of the calibrated prediction and are documented in full in the Supplementary Material.

3.2 Dataset and Preprocessing

All models are trained and evaluated on PatchCamelyon (PCam) [3], a binary tumor/normal patch classification benchmark derived from the CAMELYON16 breast-cancer lymph-node metastasis challenge [4], accessed through its public HuggingFace Datasets mirror. From the official PCam train/validation/test pools – drawn separately within each pool, not after combining them, so there is no leakage across the official splits – we draw one *fixed* random subset (seed 0, without replacement) of 20,000 training, 4,000 validation, and 8,000 test patches; the same subset indices are reused for every backbone and every training seed, so that the five training seeds (42, 123, 2024, 7, 99) vary model initialization, mini-batch order, and the stochastic training-time augmentation draws (random flips and color jitter, Section 3.2) between

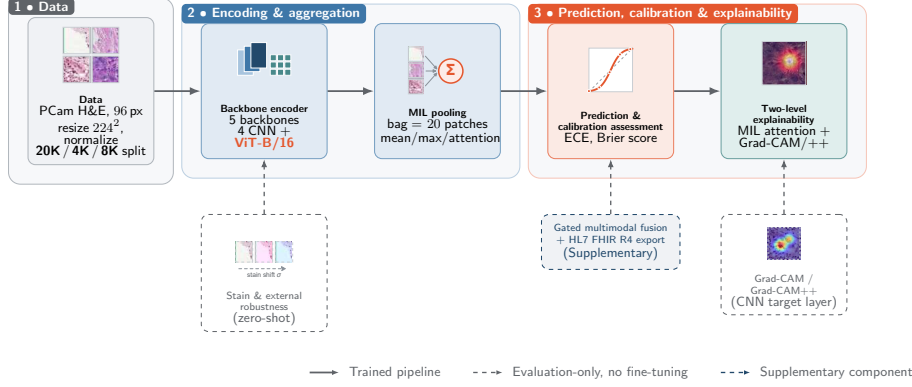


Fig. 1 Overview of the proposed framework, drawn as a left-to-right pipeline in three stages. **Stage 1 (Data)**: PCam H&E patches are resized to 224×224 and normalized, using a fixed 20,000/4,000/8,000 train/validation/test split. **Stage 2 (Encoding & aggregation)**: a patch is encoded by one of five backbones (ResNet18/50, EfficientNet-B0, DenseNet121, ViT-B/16), and synthetic bags of 20 patches are aggregated by a multiple-instance-learning pooling operator over instance embeddings h_k (mean, max, or attention pooling; Section 3.7). **Stage 3 (Prediction, calibration & explainability)**: the model emits a prediction together with a calibration assessment (ECE, Brier score) and a two-level explainability output combining bag-level MIL attention with patch-level Grad-CAM/Grad-CAM++. Dashed grey boxes are evaluation-only branches that reuse the trained backbone without fine-tuning: stain-perturbation and zero-shot external-dataset robustness testing (below the encoder) and CNN-layer Grad-CAM/Grad-CAM++ (below the explainability stage). The dashed blue box marks two deployment components – gated multimodal fusion and an HL7 FHIR R4 export stage – implemented and evaluated in full but documented with complete architectural detail in the Supplementary Material.

runs, not which patches are used. Native PCam patches are 96×96 pixels; all backbones are trained on patches resized to 224×224 and normalized per RGB channel c with ImageNet channel statistics,

$$x'_c = \frac{x_c - \mu_c}{\sigma_c}, \quad \mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225], \quad (1)$$

applied independently to each of the three channels after resizing. Training-time augmentation consists of random horizontal and vertical flips and mild color jitter (brightness and contrast factor 0.1); evaluation, robustness, and external-validation transforms apply only resize and normalization, with no augmentation.

3.3 Backbone Architectures and Training Protocol

ViT-B/16 is included specifically, rather than a larger or pathology-pretrained transformer, to separate *architecture family* (global self-attention vs. local convolution) from pretraining data scale as candidate explanations (Section 2): a plain ImageNet-only ViT lets us ask whether transformer robustness gains require pathology-specific pretraining, which a pathology-pretrained model could not disambiguate. Because ViT-B/16 also differs from the four CNN backbones in parameter count and learning rate (Section 3.3), and only one transformer architecture is evaluated, this design rules out pretraining-data scale as an explanation but does not fully isolate architecture as

a single causal factor; we report the resulting comparison as an architecture-scale association under a fixed, documented protocol (Section 7). Five backbones are compared: ResNet18, ResNet50, EfficientNet-B0, DenseNet121, and ViT-B/16, each initialized from ImageNet-1K weights (`torchvision` `IMAGENET1K_V1`) and fine-tuned end-to-end (no layers frozen) with the original classification head replaced by a linear layer mapping to two classes. Each backbone f_θ maps a patch x to class logits $f_\theta(x) \in \mathbb{R}^2$, converted to a tumor probability by the softmax,

$$p_\theta(x) = \frac{\exp(f_\theta(x)_1)}{\exp(f_\theta(x)_0) + \exp(f_\theta(x)_1)}, \quad (2)$$

and trained by minimizing the binary cross-entropy loss over a mini-batch $\mathcal{B}$ of size $|\mathcal{B}| = 128$,

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[y_i \log p_\theta(x_i) + (1 - y_i) \log (1 - p_\theta(x_i)) \right], \quad (3)$$

where $y_i \in \{0, 1\}$ is the ground-truth label of patch x_i . Each backbone is trained for 10 epochs with the Adam optimizer minimizing Eq. (3), weight decay 10^{-4} ; the learning rate is 10^{-4} for the four CNN backbones and 3×10^{-5} ($10^{-4} \times 0.3$) for ViT-B/16, reduced by a factor of 0.5 on validation-AUC plateaus (patience 2 epochs). Batch size is 128 for all backbones. Each of the five backbones is trained independently for each of the five seeds (25 trained models total), and all reported main-results metrics (Section 5) are mean $\pm$ standard deviation over the five seeds.

3.4 Stain-Perturbation Robustness Protocol

Robustness is evaluated with a custom RGB channel-gain corruption, motivated by – but not equivalent to – the stain/color-shift robustness studies of [13], which investigated explicit stain-color augmentation and normalization methods, including stain-space (e.g., hematoxylin/eosin color-deconvolution) approaches. We apply a multiplicative per-channel intensity perturbation directly to the RGB pixel values x_c of each test-set image before the standard resize-and-normalize evaluation transform, rather than perturbing in a stain-vector space; we refer to this throughout as RGB channel-gain corruption, a generic color-sensitivity proxy rather than a biophysical H&E stain-mixing model. For perturbation strength σ , each of the three color channels $c \in \{R, G, B\}$ is independently scaled by

$$\tilde{x}_c = \text{clip}(\alpha_c x_c, 0, 1), \quad \alpha_c \sim \mathcal{N}(1, \sigma^2), \quad (4)$$

with $\sigma \in \{0, 0.05, 0.10, 0.15, 0.20\}$. No perturbation is applied during training; each of the 25 trained models (5 backbones $\times$ 5 seeds) is re-evaluated on the same fixed 8,000-patch test set at each σ level, and relative AUC degradation at strength σ is

computed against each model’s own $\sigma = 0$ test AUC as

$$\Delta_{\text{AUC}}(\sigma) = \frac{\text{AUC}(\sigma = 0) - \text{AUC}(\sigma)}{\text{AUC}(\sigma = 0)} \times 100\%. \quad (5)$$

To complement this generic RGB channel-gain corruption with a stain-specific perturbation closer in spirit to [13], we additionally evaluate all 25 trained models under a hematoxylin-eosin (H&E) optical-density perturbation. Each test image is separated into hematoxylin, eosin, and DAB optical-density channels via the fixed stain-deconvolution matrix of [30], the hematoxylin and eosin channels are independently scaled by a multiplicative factor $\beta_h, \beta_e \sim \mathcal{N}(1, \sigma^2)$ (the DAB channel, uninformative for H&E-only images, is left unperturbed), and the result is converted back to RGB before the standard resize-and-normalize evaluation transform, using the same $\sigma \in \{0.05, 0.10, 0.15, 0.20\}$ grid as the RGB protocol. Unlike the RGB channel-gain corruption, this HED perturbation is constrained to the two-dimensional stain subspace the H&E staining process actually varies in, so it is a stricter test of stain robustness specifically, at the cost of a narrower range of induced AUC change than the RGB protocol (Section 5).

3.5 Calibration Metrics

For each backbone, Expected Calibration Error (ECE) is computed by partitioning predicted probabilities on the test set into 15 equal-width bins $B_i \subset (0, 1]$ and computing

$$\text{ECE} = \sum_{i=1}^{15} \frac{|B_i|}{N} |\text{acc}(B_i) - \text{conf}(B_i)|,$$

where $\text{acc}(B_i)$ and $\text{conf}(B_i)$ are the empirical accuracy and mean predicted confidence of samples falling in bin B_i , and N is the test-set size, following the calibration-measurement approach established for modern deep networks by [19]. Brier score is reported alongside ECE as a proper scoring rule,

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N (p_\theta(x_i) - y_i)^2, \quad (6)$$

where a lower Brier score indicates predicted probabilities closer to the observed binary outcomes. ECE summarizes bin-wise differences between confidence and empirical accuracy, whereas the Brier score is a proper scoring rule reflecting both probabilistic accuracy and calibration-related behavior.

3.6 Diagnostic and Discrimination Metrics

Let TP, TN, FP, FN denote true and false positives and negatives at decision threshold $\tau = 0.5$ on predicted probability $p_\theta(x)$ (Eq. (2)). From these we report the standard clinical diagnostic metrics (Table 4): accuracy, sensitivity, specificity, positive and negative predictive value (PPV, NPV), F_1 , and the Matthews correlation coefficient

(MCC), which remains informative under class imbalance – relevant for the 17.2% tumor-prevalence PathMNIST external evaluation (Section 3.12). AUC is computed threshold-independently via its Mann–Whitney U -statistic equivalence: for positive-class patches $\mathcal{P}$ and negative-class patches $\mathcal{N}$ in the test set,

$$\text{AUC} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} [\mathbf{1}(p_\theta(x_i) > p_\theta(x_j)) + \frac{1}{2} \mathbf{1}(p_\theta(x_i) = p_\theta(x_j))], \quad (7)$$

i.e. the probability that a randomly drawn positive patch receives a higher predicted probability than a randomly drawn negative patch, which is the quantity the DeLong test (Section 3.10) formally compares between two correlated classifiers.

3.7 Multiple-Instance-Learning Formulation (Summary)

We compare all four pooling operators rather than adopting attention pooling by default, because whether fixed (mean/max) or learned (attention/gated) pooling suits a given bag-construction regime is an empirical question, not a settled default (Section 6.3). Patches are grouped into synthetic bags of size 20 (PCam has no native slide grouping): 20 indices are drawn without replacement per bag, and the bag label is positive if more than one third of its patches are tumor-positive. Frozen penultimate-layer features (dimensionality 512/2048/1280/1024/768 for ResNet18/ResNet50/EfficientNet-B0/DenseNet121/ViT-B/16) serve as bag instances. All four operators are trained with Adam (learning rate 5×10^{-4} , weight decay 10^{-4}) for up to 50 epochs with early stopping (patience 10 on validation AUC): (i) mean and (ii) max pooling, each with a two-layer classifier; (iii) attention pooling [20], $a_k = \text{softmax}_k(w^\top \tanh(Vh_k))$ over instance embeddings h_k ; and (iv) gated attention, $a_k = \text{softmax}_k(w^\top (\tanh(Vh_k) \odot \text{sigm}(Uh_k)))$. In both attention variants the bag representation $\sum_k a_k h_k$ feeds a two-layer head. Mean pooling is the strongest of the four on our synthetic bags (bag AUC 0.9468 ± 0.0226), ahead of max pooling (0.9209 ± 0.0206), while both attention-based operators underperform with substantially higher seed-to-seed variance (gated attention 0.8567 ± 0.2221 ; plain attention 0.6062 ± 0.1713 , near chance on some seeds) – a result we interpret as a property of this synthetic bag construction rather than a claim about attention-based MIL in general (Section 6.3). This being a supporting ablation rather than the paper’s central claim, the full ablation table, pooling diagram, and per-seed variance are reported in full in Supplementary Material Section S1.

3.8 Multimodal Fusion (Summary)

Because PCam provides no clinical metadata, a 7-dimensional structured clinical vector is simulated per sample: patient age (1 dimension), tissue type (a 3-category one-hot encoding), stain intensity (1 dimension), lymphocytic infiltration (1 binary dimension), and biomarker score (1 dimension), together giving $1 + 3 + 1 + 1 + 1 = 7$ dimensions. Each variable is label-conditioned so that plausible values correlate with tumor status rather than being noise. This deliberately synthetic experiment is an implementation sanity check for the fusion architecture and provides no evidence

of clinical utility from real multimodal patient data; it is explicitly a framework-demonstration proxy, not a clinical dataset (Section 7). Five strategies sharing a two-branch architecture – Image-Only, Clinical-Only, Early Fusion, Late Fusion, and Gated Fusion (a learned per-sample mixing weight) – are compared; Gated Fusion, the strongest candidate because its input-conditioned gate can down-weight a locally uninformative modality, achieves the highest test AUC (0.9805 ± 0.0021 , Section 5). Full equations, the fusion-branch diagram, and the five-strategy ablation table are in Supplementary Material Section S1.

3.9 Explainability Pipeline

Patch-level explainability uses Grad-CAM [21] and Grad-CAM++ [24] on the best CNN backbone (EfficientNet-B0), with the final convolutional block as target layer (**layer4** for ResNet18/50, the final **features** block for EfficientNet-B0, the last dense layer of block 4 for DenseNet121), since gradient-based CAM does not apply directly to transformer tokens. For ViT-B/16 – the backbone identified as this paper’s strongest and most robust architecture (Sections 5, 3.4) – we instead compute attention roll-out [31]: the attention weight matrices from all 12 self-attention layers (averaged over heads) are each combined with an identity matrix to account for the residual connection, row-normalized, and multiplied across layers, and the CLS token’s resulting row over the 196 patch tokens is reshaped to the 14×14 patch grid and upsampled to pixel resolution, giving a saliency map in the same format as the CNN’s Grad-CAM output. We run Grad-CAM++ alongside Grad-CAM because its pixel-wise weighted gradients localize better when a patch contains multiple small, disjoint relevant regions (tumor-margin, micro-metastasis), letting us check whether saliency conclusions are method-specific; full weighting equations are in Supplementary Material Section S3. Faithfulness is quantified with the standard insertion/deletion protocol (20 steps, tracking the target-class softmax as saliency-ranked pixels are progressively revealed or removed, giving Insertion and Deletion AUC) and average confidence drop/increase under full-saliency masking, on the same 200 randomly sampled tumor test patches for both the CNN Grad-CAM pipeline and the ViT rollout pipeline, so the two are directly comparable (definitions in Supplementary Material Section S3). Bag-level, attention-based interpretability uses the MIL attention weights of Section 3.7 as patch-importance scores over the synthetic patch bags defined there (Section 7 discusses the synthetic-bag caveat).

3.10 Statistical Testing

Pairwise AUC differences between all $\binom{5}{2} = 10$ backbone pairs are tested two ways. (i) The DeLong test for correlated ROC curves compares two AUCs (Eq. (7)) estimated on the *same* test set by computing structural components (placement values) for each classifier and testing whether their difference $\widehat{\text{AUC}}_A - \widehat{\text{AUC}}_B$ is zero under a normal

approximation,

$$z = \frac{\widehat{\text{AUC}}_A - \widehat{\text{AUC}}_B}{\sqrt{\text{Var}(\widehat{\text{AUC}}_A) + \text{Var}(\widehat{\text{AUC}}_B) - 2 \text{Cov}(\widehat{\text{AUC}}_A, \widehat{\text{AUC}}_B)}}, \quad (8)$$

computed via the Sun–Xu placement-values formulation. Because the five per-seed models for a given backbone are not independent test-set replicates – all 25 trained models are evaluated on the identical fixed 8,000-patch test set – pooling all five seeds’ predictions into a single $N = 40,000$ DeLong test would treat five repeated evaluations of the same 8,000 samples as 40,000 independent observations, artificially shrinking the estimated variance and inflating significance. We therefore run the DeLong test separately for each of the 5 seeds (correctly paired within a seed: model A and model B evaluated on the identical 8,000 test-set predictions), Bonferroni-correct each seed’s 10 pairwise p -values independently, and summarize the resulting 5 per-seed verdicts for each backbone pair (Section 5) rather than reporting a single pooled statistic. Because correcting all $\binom{5}{2} = 10$ backbone pairs treats every comparison as equally central to the paper’s claim, we additionally pre-specify a *primary family* of four hypotheses directly tied to the paper’s central claim – ViT-B/16 vs. each of the four CNN backbones – and report these with Holm’s sequential correction, which controls the familywise error rate while being less conservative than Bonferroni; the remaining six CNN-vs-CNN pairs are treated as a secondary, exploratory family (Section 5). (ii) A paired t -test across the five per-seed AUC values for each backbone pair, evaluated against a t -distribution with 4 degrees of freedom; with only 5 seeds this test has low power, and we report it as a small- n training-stochasticity analysis rather than a definitive hypothesis test. External-validation domain shift (Section 3.12) is tested per backbone and per dataset as a *paired* run-level analysis: for each of the five training seeds, the same trained model produces both a PCam AUC and an external-dataset AUC, so we analyze the five paired per-seed drops $d_r = \text{AUC}_{\text{external},r} - \text{AUC}_{\text{PCam},r}$ with a paired t -test and 95% confidence interval, rather than treating the ten AUC values as two independent samples of five.

Four further statistics – McNemar’s test (paired per-sample decisions, applied only to same-dataset paired predictions, never across datasets), Cohen’s d and Cliff’s δ (standardized and non-parametric effect size), and decision curve analysis (clinical net benefit at a given decision threshold) – are used in Section 5 to quantify *how large* and *how decision-relevant* the significant differences established above are; their formal definitions, and the full per-domain result tables, are given in Supplementary Material Section S4. Bootstrap 95% confidence intervals for AUC and Brier score (2,000 resamples with replacement of the test predictions) are likewise defined and tabulated in full in Supplementary Material Section S4. Both the McNemar test and the patch-level bootstrap treat individual patches as independent observational units; PCam patches can originate from a shared source whole-slide image, so if multiple test patches share a source slide, these intervals and p -values may be somewhat optimistic (too narrow) relative to a slide-clustered analysis. We report this as a limitation (Section 7) rather than a cluster-robust estimate, since source-slide identifiers are not exposed by the public PCam distribution we use.

3.11 Computational Benchmarking

For each backbone, parameter count, model size, and FLOPs (via `fvcore`’s `FlopCountAnalysis` on a single 224×224 input) are recorded. GPU latency and throughput are measured on a single NVIDIA L4 GPU (Google Colab) after 10 warm-up forward passes, timed over 50 subsequent forward passes at batch size 1 and batch size 32, with `torch.cuda.synchronize()` calls bracketing each timed block; peak GPU memory is read from `torch.cuda.max_memory_allocated()` after resetting peak-memory statistics. CPU latency is measured analogously at batch size 1 on the same host’s CPU.

3.12 External Validation Datasets

Each of the 25 PCam-trained models (5 backbones $\times$ 5 seeds) is evaluated *zero-shot* – i.e., with no additional fine-tuning or parameter updates – on independent public histopathology datasets. BACH and PathMNIST differ from PCam not only in scanner and staining source but in tissue type, diagnostic task, and label definition (Table 1), so evaluation on those two datasets is best read as zero-shot transfer across dataset and task rather than a same-task, different-hospital replication. To additionally test a same-task, cross-hospital setting – which isolates site/scanner shift without the confounding tissue-type and label-definition changes that BACH and PathMNIST introduce – we evaluate on **CAMELYON17-WILDS** [32, 33], the WILDS benchmark’s patch-level release of the CAMELYON17 breast-cancer lymph-node challenge: like PCam, its task is binary tumor/non-tumor patch classification on H&E lymph-node tissue at 96×96 pixels, but the patches originate from a different, held-out hospital center (the `ood_test` split, 85,054 patches, none seen during PCam’s CAMELYON16-derived training data), so this is same-task, cross-hospital transfer rather than cross-task transfer. We evaluate on a fixed random subset of 5,000 patches (seed 42, class balance 2,544 non-tumor / 2,456 tumor). **BACH** [17] (400 breast-biopsy microscopy images, 4 classes, official images 2048×1536 pixels at $0.42 \mu\text{m}/\text{pixel}$, confirmed directly from the loaded images) is binarized as Benign+Normal $\rightarrow$ 0 and InSitu+Invasive $\rightarrow$ 1. We report BACH results under two protocols. The primary protocol, used throughout unless noted otherwise, resizes each full BACH image directly to 224×224 with no tiling or cropping, consistent with our evaluation transform for all datasets (Section 3.2); this combines stain/scanner shift with a substantial field-of-view and physical-scale shift relative to PCam’s native 96×96 patches. To disentangle these two confounded shifts, we additionally evaluate a scale-aware tiling protocol: each BACH image is partitioned into non-overlapping 224×224 tiles at native resolution (no resize, so each tile is a genuine sub-region of tissue rather than a globally rescaled full image), tiles whose grayscale mean exceeds 220 and standard deviation is below 8 are discarded as background/whitespace, and the remaining per-tile probabilities are aggregated to an image-level prediction by mean and by max pooling; both aggregators are reported, mean pooling because it was also the stronger of the two on the paper’s synthetic-bag MIL ablation (Section 3.7), max pooling as the standard alternative. This tiling ablation and its results are reported in Section 5.7. **PathMNIST**, accessed via the `medmnist` package [25] (`PathMNIST(split='test', size=224)`), is evaluated

on a fixed random subset of 5,000 images (seed 42) from its official test split with the tumor-epithelium class (class index 8) binarized against all other tissue classes. We use **PathMNIST** rather than **NCT-CRC-HE-100K** to name this dataset because the two are not interchangeable: PathMNIST’s training and validation portions derive from the NCT-CRC-HE-100K corpus [18], but its official held-out test split – the split we evaluate – derives from the separate CRC-VAL-HE-7K source, collected at a different clinical center; this distinct-center provenance is precisely what gives the test split external-domain value, so we refer to it throughout as the PathMNIST test split (derived from CRC-VAL-HE-7K) rather than as NCT-CRC-HE-100K. **MHIST** [26] (colorectal polyps) was also attempted but could not be obtained from any public dataset mirror queried at evaluation time; consistent with our policy of reporting only verified results, MHIST is excluded rather than estimated. For each successfully evaluated external dataset, AUC, accuracy, sensitivity, and specificity are computed per backbone per seed. A per-backbone cross-dataset consistency score summarizing AUC stability across the four evaluated domains (PCam, BACH, PathMNIST, CAMELYON17-WILDS) is defined and reported in full in Supplementary Material Section S4.

Table 1 summarizes the composition of all four datasets used in this study. PCam is constructed as an approximately class-balanced benchmark, and the fixed 8,000-patch test subset drawn from it is verified near-balanced (4,040 normal/3,960 tumor) directly from ground-truth labels. BACH, PathMNIST, and CAMELYON17-WILDS class counts are computed directly from the evaluation labels used throughout Section 5.

Table 1 Composition of the datasets used in this study. PCam counts for training/validation are the fixed subset sizes from Section 3.2; all other counts are verified directly from ground-truth labels.

Dataset	Role	N	Class composition
PCam [3]	Training	20,000	Approx. balanced (population-level design)
PCam [3]	Validation	4,000	Approx. balanced (population-level design)
PCam [3]	Test (in-domain)	8,000	4,040 normal / 3,960 tumor (verified)
BACH [17]	External test	400	200 benign / 200 malignant (verified)
PathMNIST [25] test split (derived from CRC-VAL-HE-7K [18])	External test	5,000	4,141 other tissue / 859 tumor-epithelium, 17.2% prevalence (verified)
CAMELYON17-WILDS [32, 33] ood_test split	External test	5,000	2,544 non-tumor / 2,456 tumor, fixed random subset (seed 42) of the 85,054-patch held-out split (verified)

4 Experiments

4.1 Research Questions

The experiments are designed to address four research objectives directly tied to the framework components described in Section 3. (RQ1) Compare CNN and transformer backbones on raw classification accuracy under a fixed, low-resource training budget. (RQ2) Determine whether this accuracy ranking changes under stain-color perturbation, and quantify the per-architecture degradation. (RQ3) Assess whether the resulting probability outputs are well calibrated, and whether calibration tracks the robustness ranking. (RQ4) Evaluate how well PCam-trained models generalize zero-shot to independent external histopathology datasets that differ in tissue, task, and label definition, and quantify the consistency of that generalization across backbones.

4.2 Experimental Design

The nine experiments (E1–E9) run to answer RQ1–RQ4, each corresponding to a Methodology and a Results subsection below, are listed in Supplementary Material Section S7.

4.3 Implementation Environment

All experiments were implemented in PyTorch/torchvision and run on Google Colab using a single NVIDIA L4 GPU per training job; `scikit-learn` and `SciPy` were used for metrics and statistical tests, and `fvcore` for FLOPs counting (Section 3.11). All 25 main-experiment models (5 backbones $\times$ 5 seeds), all MIL and fusion ablation models, and all robustness/calibration/explainability evaluations were produced by a single, fixed data pipeline (Section 3.2), so that reported differences reflect architecture and method choice rather than incidental data variation.

5 Results

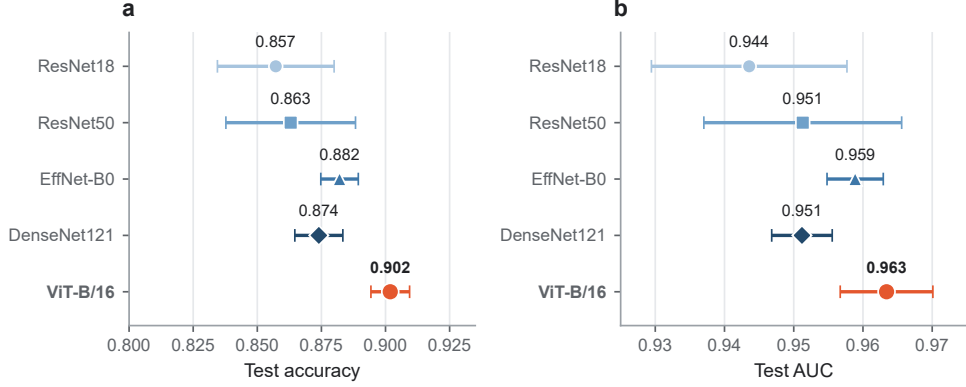
5.1 Backbone Comparison and Statistical Significance

Table 2 reports test accuracy, AUC, F1, and ECE for all five backbones, each averaged over five seeds. ViT-B/16 attains the highest mean test AUC (0.9634 ± 0.0076) and accuracy (0.9018 ± 0.0086), followed by EfficientNet-B0 (AUC 0.9589 ± 0.0046). ResNet18 is the weakest backbone on both metrics (AUC 0.9436 ± 0.0161 , accuracy 0.8572 ± 0.0260), and also shows the largest AUC variance across seeds among the CNNs. Figure 2 plots the same accuracy and AUC values with 95% confidence-interval error bars.

Table 3 reports pairwise AUC differences for all $\binom{5}{2} = 10$ backbone pairs: the mean AUC difference across the five seeds, how many of the 5 per-seed DeLong tests reach significance after Bonferroni correction (Section 3.10), the least significant (largest) of those 5 Bonferroni-corrected DeLong p -values, and the paired t -test p -value across the five seed-level AUCs. Every backbone pair reaches DeLong significance in at least 3 of 5 seeds, and DenseNet121 vs. ViT-B/16 is significant in all 5. ViT-B/16 is significantly

Table 2 Main results on PatchCamelyon. All values are mean $\pm$ std over 5 seeds.

Backbone	Accuracy	AUC	F1	ECE
ResNet18	0.8572 $\pm$ 0.0260	0.9436 $\pm$ 0.0161	0.8411 $\pm$ 0.0354	0.0939
ResNet50	0.8630 $\pm$ 0.0288	0.9513 $\pm$ 0.0163	0.8455 $\pm$ 0.0398	0.0913
EfficientNet-B0	0.8821 $\pm$ 0.0083	0.9589 $\pm$ 0.0046	0.8712 $\pm$ 0.0107	0.0710
DenseNet121	0.8740 $\pm$ 0.0107	0.9512 $\pm$ 0.0050	0.8621 $\pm$ 0.0158	0.0770
ViT-B/16	0.9018 $\pm$ 0.0086	0.9634 $\pm$ 0.0076	0.8982 $\pm$ 0.0102	0.0444

**Fig. 2** Test accuracy (a) and AUC (b) per backbone, mean $\pm$ 95% CI over 5 seeds. Values match Table 2.

better than each of the four CNN backbones by the per-seed DeLong test in 3–5 of 5 seeds, consistent in direction with its higher mean AUC in Table 2, but not uniformly significant at every individual seed: per-seed training variance is large enough that a single training run’s AUC ranking is not always statistically distinguishable from a competing backbone, even when the 5-seed aggregate ranking is stable. The paired t -test, which treats the five seeds as five independent unit observations rather than as five separate 8,000-sample test sets, finds no pair significant after correction; with only 5 seed-level AUC values per backbone this test has low statistical power. We report both tests, each computed with respect to what does and does not constitute an independent observation in our design (Section 3.10), rather than pooling seeds into a single inflated-power statistic or selecting whichever test supports a stronger claim.

Because the paper’s central claim concerns ViT-B/16 relative to the CNN backbones specifically, we additionally restrict the paired t -test to the pre-specified primary family of four contrasts (Section 3.10) and apply Holm’s correction within that smaller family rather than Bonferroni-correcting all 10 pairs together. On this primary-family analysis, ViT-B/16’s mean AUC advantage is significant against DenseNet121 (mean Δ AUC +0.0122, paired $t(4) = 5.14$, Holm-adjusted $p = 0.027$) but not against ResNet18 (+0.0198, $p = 0.089$, Holm $p = 0.269$), ResNet50 (+0.0121, $p = 0.309$,

Holm $p = 0.617$), or EfficientNet-B0 (+0.0045, $p = 0.378$, Holm $p = 0.617$). With only 5 seeds this remains a low-power test, and the clean-AUC family is secondary to the robustness result (Section 3.4) in our central claim; we report it as a small- n training-stochasticity analysis rather than a definitive hypothesis test (Section 3.10).

Table 3 Pairwise AUC differences between backbones. ΔAUC is the mean of Model A – Model B over the 5 seeds (Table 2). “DeLong sig.” is the number of seeds (of 5) at which a per-seed DeLong test – computed on paired, same-seed predictions on the fixed 8,000-sample test set, not pooled across seeds – reaches significance after Bonferroni correction (10 comparisons); “DeLong max p ” is the largest (least significant) of those 5 Bonferroni-corrected p -values. Paired- t p is Bonferroni-corrected across the 5 per-seed AUC values.

Model A	Model B	ΔAUC	DeLong sig.	DeLong max p	Paired- t p
ResNet18	ResNet50	-0.0077	4/5	0.289	1.000
ResNet18	EfficientNet-B0	-0.0153	4/5	1.000	0.942
ResNet18	DenseNet121	-0.0076	4/5	0.078	1.000
ResNet18	ViT-B/16	-0.0198	3/5	1.000	0.896
ResNet50	EfficientNet-B0	-0.0076	4/5	1.000	1.000
ResNet50	DenseNet121	+0.0001	3/5	1.000	1.000
ResNet50	ViT-B/16	-0.0121	4/5	1.000	1.000
EfficientNet-B0	DenseNet121	+0.0077	3/5	1.000	0.906
EfficientNet-B0	ViT-B/16	-0.0045	4/5	0.320	1.000
DenseNet121	ViT-B/16	-0.0122	5/5	0.005	0.068

5.2 Calibration

Table 2 and Figure 3(a) report ECE per backbone (5-seed mean, with 95% confidence intervals). ViT-B/16 has the lowest ECE (0.0444), a 37.5% relative improvement over the best-calibrated CNN, EfficientNet-B0 (0.0710), and roughly half the ECE of the two ResNet variants (0.0913–0.0939). Figure 3(b) plots ECE against Brier score, showing ViT-B/16 with the lowest (best) values on both calibration axes in-domain; combined with its highest in-domain AUC (Table 2), ViT-B/16 is jointly the best-discriminating and best-calibrated backbone in-domain, though ECE/Brier alone establish calibration, not discrimination. Brier scores follow a broadly similar ranking in-domain, though the two measures diverge under domain shift (Section 5.8).

5.3 Stain-Perturbation Robustness

Figure 4 shows test AUC as a function of RGB channel-gain corruption strength σ for all five backbones. At the strongest corruption tested ($\sigma = 0.20$), ViT-B/16’s AUC falls from 0.9634 to 0.9488, a relative drop of 1.51%. The four CNN backbones degrade substantially more over the same range: ResNet18 drops 8.17% (0.9436 $\rightarrow$ 0.8665), ResNet50 drops 10.99% (0.9513 $\rightarrow$ 0.8468), EfficientNet-B0 drops 8.00% (0.9589 $\rightarrow$ 0.8821), and DenseNet121 drops 7.66% (0.9512 $\rightarrow$ 0.8784). ViT-B/16 therefore suffers roughly one-fifth to one-seventh of the relative degradation of the CNNs.

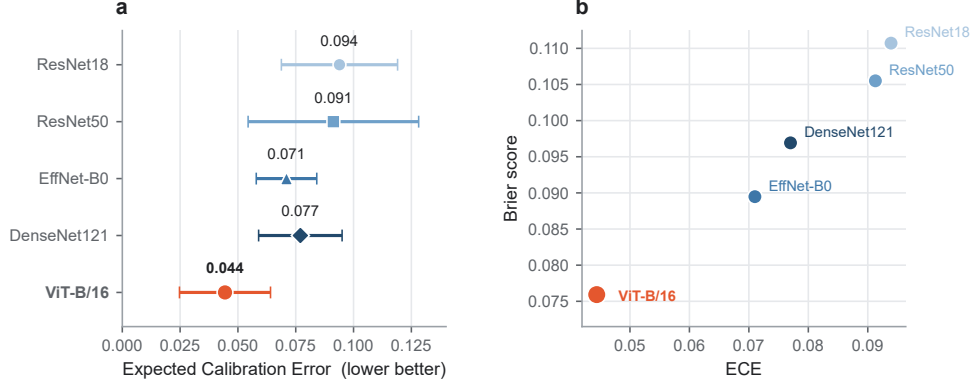


Fig. 3 Calibration on PCam (5-seed means). (a) Expected Calibration Error per backbone with 95% confidence intervals (lower is better); ViT-B/16 has the lowest mean ECE and Brier score in-domain, at 0.044. (b) ECE versus Brier score: ViT-B/16 is lowest (best) on both calibration axes in-domain; AUC (Table 2, Figure 2), not this panel, establishes that it is also the best-discriminating backbone.

We additionally test this directly with the estimand the claim actually concerns: the per-seed, per-severity drop $D_{m,\sigma} = \text{AUC}_{m,0} - \text{AUC}_{m,\sigma}$ and the primary contrast $\Delta D_{m,\sigma} = D_{\text{ViT},\sigma} - D_{\text{CNN}_{m,\sigma}}$, computed per seed (matched by training seed across backbones, since a given seed’s corruption realizations are drawn from the same restarted random state for every backbone) and tested with a paired t -test. At $\sigma = 0.20$, the primary family of four contrasts, Holm-corrected, is uniformly significant: ViT vs. ResNet18 $\Delta D = -0.0624$ (95% CI -0.084 to -0.041 , $t(4) = -8.08$, Holm $p = 0.0029$); ViT vs. ResNet50 $\Delta D = -0.0899$ (CI -0.118 to -0.061 , $t = -8.76$, Holm $p = 0.0029$); ViT vs. EfficientNet-B0 $\Delta D = -0.0621$ (CI -0.089 to -0.035 , $t = -6.32$, Holm $p = 0.0032$); ViT vs. DenseNet121 $\Delta D = -0.0582$ (CI -0.076 to -0.041 , $t = -9.32$, Holm $p = 0.0029$). Repeating this at every tested severity ($\sigma = 0.05, 0.10, 0.15, 0.20$) gives the same qualitative pattern throughout, with ΔD growing monotonically more negative as σ increases (e.g., ViT vs. ResNet50: -0.0156 at $\sigma = 0.05$ up to -0.0899 at $\sigma = 0.20$, all four contrasts significant at every severity level, Holm-corrected $p < 0.005$ throughout) – an architecture $\times$ severity interaction in the informal sense that the ViT/CNN degradation gap widens with corruption strength, not only a difference at the single most-corrupted setting. This confirmatory analysis, together with the descriptive comparison above, is the central empirical result of this paper (Section 6).

Corruption-instance variance is negligible relative to training-seed variance. Because our evaluation code draws a single corruption realization per (backbone, seed) pair (Section 3.4), the reported per-seed variability at each σ could in principle reflect which random corruption pixels were drawn rather than genuine training-seed variability. We checked this directly at $\sigma = 0.20$: 10 independently-seeded, fixed corruption replicates, each evaluated by all 25 trained models, give 50 AUC values per backbone that decompose into variance across the 5 training seeds (averaged over replicates)

and variance across the 10 replicates (averaged over seeds). For every backbone, replicate variance is 2–3 orders of magnitude smaller than seed variance (e.g., ViT-B/16: 1.06×10^{-4} across seeds versus 4×10^{-7} across replicates; DenseNet121, the largest ratio observed, is still only 1% of seed variance). The 5-seed, single-realization confidence intervals reported throughout this paper are therefore not meaningfully inflated or biased by corruption-draw luck; essentially all of the reported spread reflects training stochasticity, not which corrupted pixels a given evaluation happened to see.

Confirmation under a second, stain-specific perturbation. Because the RGB channel-gain corruption above is a generic color-sensitivity proxy rather than a biophysical stain model (Section 3.4), we repeated the full evaluation under the HED optical-density perturbation of Section 3.4, which is constrained to the hematoxylin/eosin stain subspace. This perturbation is substantially gentler than the RGB corruption – it induces a much smaller absolute AUC change at every σ , since it cannot alter image properties (brightness, non-stain color balance) outside the two-dimensional stain subspace – but the same ranking holds. At $\sigma = 0.20$, relative AUC degradation is 0.37% (ResNet18), 0.76% (ResNet50), 0.61% (EfficientNet-B0), 0.43% (DenseNet121), and 0.10% (ViT-B/16): ViT-B/16 remains the least-degraded backbone by a factor of 3.5–7.3 $\times$ relative to the four CNNs, the same multiplicative range as under the RGB corruption. The paired, Holm-corrected primary-family test at $\sigma = 0.20$ is significant for three of the four contrasts at this smaller effect size (ViT vs. ResNet50 $\Delta D = -0.0061$, CI -0.0088 to -0.0034 , Holm $p = 0.0122$; ViT vs. EfficientNet-B0 $\Delta D = -0.0048$, CI -0.0068 to -0.0027 , Holm $p = 0.0122$; ViT vs. DenseNet121 $\Delta D = -0.0031$, CI -0.0046 to -0.0016 , Holm $p = 0.0122$) but not for ViT vs. ResNet18 ($\Delta D = -0.0024$, CI -0.0071 to $+0.0022$, Holm $p = 0.215$), where the smaller absolute degradation under this gentler perturbation leaves less room for a 5-seed paired test to resolve a difference. We read this as the robustness finding replicating under a second, more stain-specific perturbation mechanism, with the appropriately weaker statistical signal that a smaller true effect and an unchanged $n = 5$ predicts, rather than as evidence the finding is perturbation-specific.

5.4 Clinical Diagnostic Metrics

Table 4 reports sensitivity, specificity, PPV, NPV, and MCC per backbone. ViT-B/16 has the highest sensitivity (0.8755) and MCC (0.8051) of all five backbones, at a modest cost in specificity (0.9277, the lowest of the five) relative to ResNet50 (0.9590, the highest). All five backbones use the same fixed decision threshold ($\tau = 0.5$, Section 3.6), so this sensitivity/specificity difference reflects where each backbone’s learned probability distribution falls relative to that common threshold, not a difference in threshold choice. At the common threshold of 0.5, ViT-B/16 operated at a higher-sensitivity, lower-specificity point than the evaluated CNNs. Whether this operating point is preferable depends on the intended clinical workflow – in a screening context, false negatives (missed tumor) are often considered costlier than false positives (which trigger secondary review) – and was not evaluated prospectively in this patch-level benchmark. Detailed in-domain confusion matrices and representative failure cases are given in Supplementary Material Section S5.

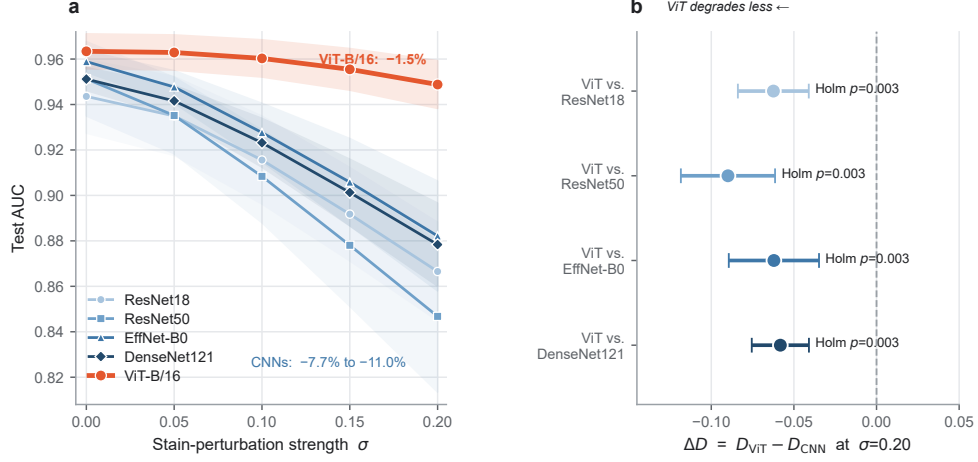


Fig. 4 (a) Test AUC vs. RGB channel-gain corruption strength σ , mean $\pm$ std over 5 seeds per backbone. ViT-B/16 exhibits substantially smaller mean AUC degradation across the tested corruption levels than the four CNN backbones, which degrade steadily. (b) Forest plot of the paired, seed-matched difference-in-degradation $\Delta D = D_{ViT} - D_{CNN}$ at $\sigma = 0.20$ for each of the four primary ViT-vs-CNN contrasts, with 95% confidence intervals and Holm-corrected p -values (Section 3.4); all four intervals exclude zero.

Table 4 Clinical metrics per backbone (5-seed mean).

Backbone	Sen	Spec	PPV	NPV	MCC
ResNet18	0.7714	0.9413	0.9308	0.8113	0.7272
ResNet50	0.7652	0.9590	0.9491	0.8087	0.7406
EfficientNet-B0	0.8063	0.9564	0.9478	0.8345	0.7724
DenseNet121	0.7983	0.9482	0.9390	0.8286	0.7570
ViT-B/16	0.8755	0.9277	0.9228	0.8843	0.8051

5.5 Computational Cost

Table 5 reports parameter count, FLOPs, GPU/CPU latency, throughput, peak memory, training time, and on-disk model size for all five backbones, measured on a single NVIDIA L4 GPU (Google Colab). EfficientNet-B0 is the smallest and most FLOP-efficient backbone (4.0M parameters, 0.40 GFLOPs, 15.3 MB) and the best accuracy-per-FLOP trade-off overall (Figure 5). ResNet18 has the lowest single-image latency on both GPU (2.75 ms, batch 1) and CPU (16.1 ms), reflecting its shallow, low-parameter design. ViT-B/16 (5.54 ms GPU, batch 1) is competitive with, but not faster than, ResNet18 for single-image inference, while showing the worst batch-32 throughput (479 FPS, 6–18 $\times$ slower than the CNNs), the highest peak GPU memory (667 MB), the highest CPU latency (132.9 ms), the largest on-disk size (327.4 MB), and the longest training time (52.8 min for 10 epochs on 20K patches, 2.1–3.8 $\times$ longer

than the CNNs). This is consistent with self-attention’s cost profile: ViT-B/16 is the most expensive backbone to run at batch, to store, and to (re-)train on this hardware, and its single-image latency, while reasonable, is not the most favorable among the five backbones tested.

Table 5 Computational benchmarking per backbone, measured on a single NVIDIA L4 GPU (Google Colab; inference latency/throughput/memory over 10 warm-up + 50 timed forward passes; training time is one full 10-epoch run on the 20K-patch training set; model size assumes float32 on-disk storage. Latencies in ms; FPS = throughput at batch 32.)

Backbone	Params(M)	FLOPs(G)	GPU $b=1$	GPU $b=32$	FPS	CPU $b=1$	Mem (MB)
ResNet18	11.18	1.82	2.75	3.76	8514	16.1	341.6
ResNet50	23.51	4.11	6.99	11.58	2762	46.8	499.8
EfficientNet-B0	4.01	0.40	9.28	9.45	3386	22.2	425.1
DenseNet121	6.96	2.86	18.24	17.32	1848	60.4	421.2
ViT-B/16	85.80	11.29	5.54	66.85	479	132.9	666.8

Backbone	Train time (min)	Model size (MB)
ResNet18	13.9	42.7
ResNet50	23.0	89.7
EfficientNet-B0	18.4	15.3
DenseNet121	25.3	26.6
ViT-B/16	52.8	327.4

Training-set size was varied from 2,000 to 20,000 patches (3 seeds per size) to check whether the 20K-patch budget is accuracy-limiting: ViT-B/16 test AUC increases from 0.9499 at 2,000 patches to 0.9587 at 10,000, then plateaus (0.9557 at 20,000, within the confidence interval of the 10,000-patch result), indicating the training set is at or past the point of diminishing returns for this task and architecture (learning-curve figure in Supplementary Material Section S6).

The MIL pooling ablation (mean pooling strongest at bag AUC 0.9468 ± 0.0226 ; both attention variants underperforming with high seed-to-seed variance) is summarized in Section 3.7 and reported in full, with the pooling diagram and per-seed variance, in Supplementary Material Section S1; the full multimodal fusion ablation (five strategies) is likewise reported there.

5.6 Explainability Faithfulness

Grad-CAM and Grad-CAM++ were evaluated on EfficientNet-B0 (the strongest CNN backbone) over 200 randomly sampled tumor test patches. Insertion AUC is 0.749 ± 0.256 and deletion AUC is 0.316 ± 0.241 (a large insertion-minus-deletion gap indicates saliency maps that are informative about which pixels drive the prediction). Average confidence drop under full-saliency masking is 55.6%, and average confidence increase is 5.0%. These faithfulness numbers are broadly consistent with, rather than exceptional relative to, published pathology explainability results (Section 6); we report them quantitatively rather than presenting saliency maps only qualitatively, consistent with our positioning in Section 2.

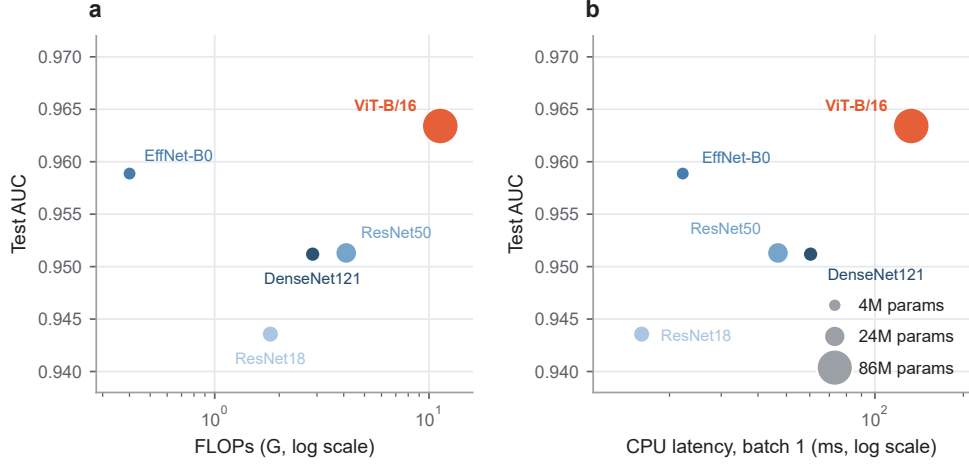


Fig. 5 Accuracy-efficiency trade-off. Mean PCam AUC is plotted against FLOPs and single-image CPU latency, with marker area proportional to parameter count. EfficientNet-B0 provides a strong AUC-per-compute trade-off (lowest FLOPs), whereas ViT-B/16 achieves the highest mean AUC at substantially greater parameter and compute cost; ResNet18, not shown as a frontier point here, has the lowest single-image CPU latency of the five backbones (Table 5).

Attention rollout on ViT-B/16 (seed 42), evaluated under the identical insertion/deletion protocol on the same 200 patches, is *less* faithful by every metric: insertion AUC 0.697 ± 0.251 (vs. 0.749), deletion AUC 0.495 ± 0.283 (vs. 0.316 – higher is worse here, since it means removing rollout-identified pixels leaves the prediction less disturbed), average confidence drop 20.0% (vs. 55.6%), and average confidence increase 15.0% (vs. 5.0%). We report this directly rather than omitting it: plain attention rollout averages relevance across all 12 self-attention layers without using gradient information, which is known to produce more diffuse, less class-discriminative maps than gradient-based CNN saliency [31], and is precisely the limitation that motivated gradient-informed transformer relevance-propagation methods such as [29]. We discuss this comparison further as an unexpected finding in Section 6.4. Bag-level, attention-based interpretability uses the MIL attention weights from Section 3.7 directly; given the MIL ablation result above, we discuss the interpretive caveats of attention-weight explainability in Section 6.

5.7 External Validation and Clinical Consistency

Table 6 reports zero-shot AUC on three independent external datasets alongside the source-domain PCam AUC. MHIST could not be obtained from any queried public mirror at evaluation time and is omitted rather than estimated (Section 3.12). Cross-dataset AUC drops are substantial for every backbone: on BACH, AUC falls to 0.32–0.63 (i.e., near or, for EfficientNet-B0, clearly below chance), and on PathMNIST to 0.83–0.90. All PCam-to-external AUC drops are statistically significant under a paired seed-level analysis (Section 3.10): for each backbone, the five per-seed drops

$d_r = \text{AUC}_{\text{external},r} - \text{AUC}_{\text{PCam},r}$ are uniformly negative, with 95% confidence intervals excluding zero for every backbone on both datasets (BACH: mean drop -0.32 to -0.64 AUC across backbones, paired t $p < 0.001$ for four backbones and $p = 0.0010$ for the fifth; PathMNIST: mean drop -0.06 to -0.13 , paired t $p \leq 0.01$ for four backbones and $p = 0.0104$ for ResNet18) – domain shift, not sampling noise. Notably, EfficientNet-B0, the strongest CNN on both PCam and stain-robustness among CNNs, generalizes worst to BACH by a wide margin (AUC 0.3216, clearly below random guessing), while ResNet50 and DenseNet121 generalize best on BACH (AUC 0.6267 and 0.6197); ViT-B/16 (0.5510), despite its in-domain and stain-perturbation advantages, is the second-worst generalizer on BACH, ahead of only EfficientNet-B0. In-domain accuracy, and even in-domain stain-perturbation robustness, is therefore not a reliable predictor of cross-dataset, cross-task generalization in our experiments – a finding that now holds most strongly for the backbone that wins on every in-domain metric, which we discuss further in Section 6.

Table 6 External validation: zero-shot AUC (5-seed mean $\pm$ std) on the source domain (PCam) and three external datasets.

Backbone	PCam AUC	BACH AUC	PathMNIST AUC	CAM17-WILDS AUC
ResNet18	0.9436 $\pm$ 0.0161	0.5730 $\pm$ 0.0335	0.8765 $\pm$ 0.0268	0.9885 $\pm$ 0.0025
ResNet50	0.9513 $\pm$ 0.0163	0.6267 $\pm$ 0.0883	0.8257 $\pm$ 0.0205	0.9897 $\pm$ 0.0038
EfficientNet-B0	0.9589 $\pm$ 0.0046	0.3216 $\pm$ 0.0631	0.8275 $\pm$ 0.0385	0.9903 $\pm$ 0.0024
DenseNet121	0.9512 $\pm$ 0.0050	0.6197 $\pm$ 0.0413	0.8802 $\pm$ 0.0053	0.9900 $\pm$ 0.0027
ViT-B/16	0.9634 $\pm$ 0.0076	0.5510 $\pm$ 0.0242	0.9038 $\pm$ 0.0231	0.9919 $\pm$ 0.0010

Unlike BACH and PathMNIST, every backbone’s AUC on CAMELYON17-WILDS *exceeds* its PCam AUC (mean increase 0.028–0.045, paired t -test $p \leq 0.003$ for all five backbones), consistent with this being a same-task, cross-hospital transfer rather than the cross-task transfer that BACH and PathMNIST represent: the domain shift is real (a genuinely held-out hospital center) but does not include the tissue-type or label-definition change that make BACH and PathMNIST harder. All five backbones cluster near ceiling (AUC 0.988–0.992), and the four primary ViT-vs-CNN contrasts are not significant here after Holm correction (all Holm $p > 0.2$) – at this AUC range there is little room for architecture to matter, and a 5-seed test lacks the power to resolve millipoint differences. We read CAMELYON17-WILDS as evidence that this paper’s robustness and calibration advantages for ViT-B/16 are most clearly established under stain/color perturbation (Section 3.4) and under cross-task transfer (BACH, PathMNIST), not as a general claim that architecture matters under any domain shift whatsoever; under a same-task, cross-hospital shift with no tissue-type or label-definition change, all five backbones generalize comparably well.

A per-backbone cross-dataset consistency score summarizing how stable each backbone’s AUC is across the four domains (PCam, BACH, PathMNIST, CAMELYON17-WILDS) is reported in full, with its formal definition, in Supplementary Material Section S4. ResNet50 has the highest score, followed by DenseNet121 and ResNet18; ViT-B/16 is second-lowest, and EfficientNet-B0 has the lowest, consistent with its poor BACH generalization above.

Scale-aware tiling substantially changes the BACH picture. Table 7 compares the full-image-resize protocol above against the scale-aware tiling ablation (Section 3.12), 5-seed mean AUC for both tile-aggregation strategies. Mean-pooled tiling improves every backbone’s BACH AUC relative to full-image-resize – consistent with mean pooling also being the stronger MIL aggregator in this paper’s synthetic-bag ablation (Section 3.7) – and the improvement is not uniform across architectures: ViT-B/16 gains the most ($0.551 \rightarrow 0.812$, a 0.261 absolute AUC increase, with a notably tight 5-seed spread, ± 0.003) and becomes unambiguously the best BACH generalizer of the five backbones, while EfficientNet-B0 also recovers substantially ($0.322 \rightarrow 0.669$) and is no longer the worst generalizer under this protocol – ResNet50 is (0.636). Max pooling recovers less of the gap than mean pooling for every backbone. We read this as evidence that a meaningful share of the full-image-resize protocol’s apparent cross-architecture differences, and specifically EfficientNet-B0’s apparent collapse, reflects the extreme field-of-view distortion of that protocol rather than a stable, architecture-intrinsic property that survives a more appropriate evaluation design; we discuss the implications for the failure-mode analysis immediately below.

Table 7 BACH AUC (5-seed mean $\pm$ std) under the full-image-resize protocol versus the scale-aware tiling ablation (mean- and max-pooled tile aggregation).

Backbone	Full-image-resize	Tiled (mean pool)	Tiled (max pool)
ResNet18	0.5730 $\pm$ 0.0335	0.7256 $\pm$ 0.0245	0.5840 $\pm$ 0.0343
ResNet50	0.6267 $\pm$ 0.0883	0.6359 $\pm$ 0.0471	0.5463 $\pm$ 0.0762
EfficientNet-B0	0.3216 $\pm$ 0.0631	0.6693 $\pm$ 0.0352	0.6066 $\pm$ 0.0393
DenseNet121	0.6197 $\pm$ 0.0413	0.6923 $\pm$ 0.0430	0.5792 $\pm$ 0.0891
ViT-B/16	0.5510 $\pm$ 0.0242	0.8123 $\pm$ 0.0032	0.6901 $\pm$ 0.0534

A degenerate external-dataset failure mode. Table 8 and Figure 6 show the seed-42 confusion matrices on BACH for all five backbones; we focus here on the two with the most divergent failure behavior. EfficientNet-B0 – the strongest CNN in-domain (Table 2) – predicts *every one* of the 400 BACH samples as positive at the $\tau = 0.5$ decision threshold (0 true negatives, 0 false negatives). That is, for the illustrated seed, EfficientNet-B0 produced all-positive binary decisions at $\tau = 0.5$, accompanied by poor ranking discrimination on BACH (AUC 0.241 for this seed, well below chance); a below-chance AUC implies the underlying predicted probabilities are not literally constant, so we describe this as a thresholded-decision collapse under the BACH color/staining distribution rather than degrading gracefully, and not as a literally constant probability output. DenseNet121, by contrast, does not collapse but produces a large number of extremely high-confidence false positives on benign tissue (predicted probability ≥ 0.9997). We were not able to visually inspect the underlying BACH tissue patches for these specific cases (Section 7), so we offer a hypothesis rather than an explanation: both EfficientNet-B0 and DenseNet121 use batch normalization layers whose running statistics are fit to PCam’s color and intensity distribution, and under a shifted stain distribution such as BACH’s, normalized activations could in principle

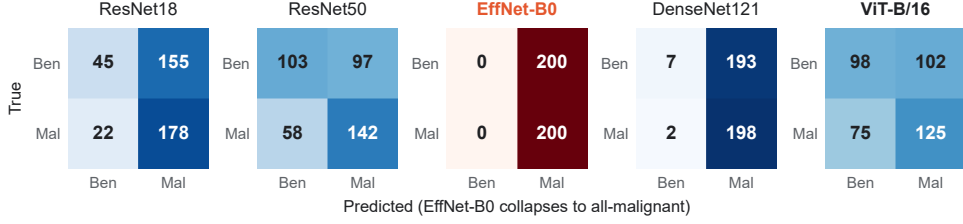


Fig. 6 Confusion matrices on the external BACH dataset ($n = 400$) for (a) ResNet18, (b) ResNet50, (c) EfficientNet-B0, (d) DenseNet121, and (e) ViT-B/16, seed-42 models, matching Table 8. Panel (c) makes EfficientNet-B0’s collapse to all-positive thresholded predictions at $\tau = 0.5$ for seed 42 directly visible: every cell in its left (Benign) column is 0.

fall outside the range the classification head was trained on, pushing a confidently-trained decision boundary to extrapolate with high confidence in the wrong direction. ViT-B/16 uses layer normalization over patch-token embeddings rather than batch normalization over spatial feature maps and has no local convolutional receptive field tuned to low-level color/texture statistics – a mechanistic difference that is directionally consistent with both its stain-perturbation robustness (Section 3.4) and its comparatively lower-confidence failure mode here, but we did not run a controlled experiment (e.g., substituting layer normalization into a CNN, or batch normalization into ViT, and re-measuring collapse) that would be needed to confirm normalization choice as the causal factor; we flag this as a specific, testable hypothesis for future work (Section 7) rather than a demonstrated mechanism. This normalization hypothesis should, however, be read alongside the tiling ablation immediately above: EfficientNet-B0’s all-positive collapse is specific to the full-image-resize protocol and substantially recovers under scale-aware tiling (BACH AUC $0.322 \rightarrow 0.669$ mean-pooled), so batch-normalization sensitivity to BACH’s stain distribution is at most a partial explanation, and the extreme field-of-view distortion of resizing a full 2048×1536 image to 224×224 is a large part of what drove the collapse under the primary protocol. Full failure-case tables (top-10/12 highest-confidence misclassifications, in-domain and external) are given in Supplementary Material Section S5.

Table 8 Confusion matrix on BACH ($n=400$, 200 positive/200 negative), seed-42 models, threshold 0.5.

Backbone	TN	FP	FN	TP
EfficientNet-B0	0	200	0	200
DenseNet121	7	193	2	198
ResNet18	45	155	22	178
ResNet50	103	97	58	142
ViT-B/16	98	102	75	125

5.8 Extended Statistical Validation

The significance testing above establishes *that* backbone differences and domain-shift drops are statistically significant. We complement this with bootstrap confidence intervals, McNemar’s test, effect sizes, and decision curve analysis (Section 3.10), using the canonical seed-42 model per backbone, applied to the in-domain PCam test set and to the two cross-task external test sets (BACH, PathMNIST); full per-domain tables and narrative, and a headline-summary table, are in Supplementary Material Section S4. CAMELYON17-WILDS (Section 5.7) is not extended into this bootstrap/McNemar/decision-curve battery: its primary finding is a null result (no significant ViT-vs-CNN difference after Holm correction), for which the paired seed-level test already reported is sufficient. The headline results are as follows. Bootstrap 95% CIs for AUC are tight and non-overlapping between the strongest and weakest backbones in-domain, corroborating the DeLong result with an independent, resampling-based method. McNemar’s test confirms that ViT-B/16’s paired per-sample decisions differ significantly from most CNN backbones in-domain and on PathMNIST, but not on BACH, where only the ViT-vs-ResNet50 pair reaches significance ($p = 0.040$) – consistent with ViT-B/16 no longer being among the strongest BACH generalizers (Section 5.7). Effect sizes for the PCam-to-external AUC drop are uniformly very large (Cohen’s d 3.0–23.0, Cliff’s $\delta = 1.0$ for every backbone on both datasets, i.e. complete stochastic dominance of in-domain over external AUC). One result runs against the calibration narrative of Section 5: on the PathMNIST test split, ViT-B/16 has both the best AUC (0.907) *and* the worst Brier score (0.192) of the five backbones, with DenseNet121 achieving the best external Brier score (0.141) despite a lower AUC – discrimination and calibration are formally independent properties and can diverge under domain shift even when they agree in-domain (Section 6.4). Decision curve analysis shows every backbone provides substantial positive net benefit in-domain, while on BACH, EfficientNet-B0 provides *no* net benefit over treat-all at any threshold tested, a direct consequence of its degenerate collapse documented above.

6 Discussion

6.1 Clinical Implications and Significance

The central practical finding is that, under identical data, budget, and protocol, ViT-B/16 combines the smallest robustness drop, the best calibration, and the highest sensitivity and MCC of the five backbones (Sections 3.4, 3.5; Table 4). In a multi-site deployment, where scanner hardware, staining reagents, and laboratory protocols vary in ways a single-site validation study cannot capture, this combination of robustness and calibration is more load-bearing for properties relevant to future clinical validation than an additional AUC point: a model whose accuracy degrades gracefully and whose confidence scores stay trustworthy under distribution shift is easier to calibrate trust around than an accurate-but-brittle one.

This matters for workflow. A classifier is only as useful as a second opinion if its stated confidence is informative; the ECE gap we observe is the difference between a

probability that is informative enough to act on and one that is not. EfficientNet-B0’s collapse to an all-positive classifier on BACH illustrates the failure: decision curve analysis (Section 5.8) shows it would flag every case for review – net benefit identical to a blanket “review everything” policy, which provides no discriminative value. The representative failure cases (Supplementary Material Section S5) are unanimously high-confidence false positives (benign called malignant) rather than false negatives, a specific, bounded harm – unnecessary follow-up rather than missed cancer – visible only because we report confusion matrices and individual cases, not AUC alone. Together, robustness, calibration, and the two-level explainability output give three independent signals for how much to trust a prediction, which we argue is more informative for downstream decision-making than point-estimate accuracy alone.

The external-validation results reinforce this: in-domain accuracy – and even in-domain stain robustness – did not predict cross-domain generalization, and a deployment decision based on PCam performance alone would have selected the worst-generalizing model (Section 5.7). This is an argument for external validation as a mandatory, not optional, part of any pathology-AI deployment pipeline.

6.2 Comparison with Modern State-of-the-Art

Table 9 places our result in the context of the PatchCamelyon literature. Our ViT-B/16 (AUC 0.9634) is competitive with the original PCam benchmark result of [3] (0.963, a rotation-equivariant CNN trained on the full 262K-patch training set), despite using only 20,000 training patches (8% of the full set) and generic ImageNet-1K pretraining. We do not describe this as matching or exceeding state-of-the-art: pathology-pretrained transformer encoders such as CTransPath [9] report substantially higher accuracy on PCam-adjacent benchmarks (~ 0.98 AUC), and foundation models such as UNI [10] and Prov-GigaPath [11] are built at a data and parameter scale roughly three to four orders of magnitude beyond what we use here; we do not report a directly comparable fine-tuned PCam number for UNI or Prov-GigaPath, since neither original paper evaluates one via full fine-tuning, and we chose not to assert a number we could not independently verify.

Table 9 Comparison with PatchCamelyon-related literature. AUC values rounded to 3 decimal places for consistency across sources; N/R = not reported by the cited source for a directly comparable fine-tuned PCam number.

Method	Venue	AUC	Params	Pretraining data
[3] (P4M-DenseNet)	MICCAI 2018	0.963	0.9M	None (full 262K PCam)
CTransPath [9]	MIA 2022	~ 0.980	27.5M	15M patches (self-sup.)
UNI [10]	Nat Med 2024	N/R	~ 307 M	100M+ patches, 100K+ slides
Prov-GigaPath [11]	Nature 2024	N/R	~ 1.1 B	1.3B patches, 171K slides
Our ViT-B/16 (this work)	–	0.963	86M	ImageNet-1K only, 20K PCam patches

This gap is expected and, we argue, not the relevant comparison for the deployment setting this paper targets: foundation models at the UNI/Prov-GigaPath scale require

institutional-scale slide archives and multi-GPU/TPU training infrastructure inaccessible to most hospital pathology departments outside a small number of well-funded centers. Table 9 compares accuracy and pretraining scale only, which understates what this paper contributes; Table 10 instead compares the evaluation *dimensions* each work reports, since accuracy-only comparison is precisely the framing we argue is insufficient for deployment decisions. No comparison work reports stain-perturbation robustness, calibration (ECE/Brier), or a standards-based interoperability export alongside its accuracy figure – including the ones with substantially higher raw accuracy than ours.

Table 10 Evaluation-dimension comparison with PatchCamelyon-related and pathology foundation-model literature. ✓ = reported in the cited primary source; ✗ = not identified in the primary publication/materials reviewed; ~ = partial/indirect coverage.

Method	Stain robustness	External val.	Calibration	Explainability (quantified)
[3]	✗	✗	✗	✗
CTransPath [9]	✗	~ (few tasks)	✗	✗
UNI [10]	✗	✓ (34 tasks)	✗	✗
Prov-GigaPath [11]	✗	✓ (broad, multi-cancer)	✗	✗
Our framework (this work)	✓ (quantified, $\sigma = 0-0.20$)	✓ (3 datasets)	✓ (ECE, Brier, bootstrap CI)	✓ (attention + Grad-CAM/++, quantified)

6.3 Interpreting the Ablation and Explainability Results

The MIL ablation result (Section 3.7; full table in Supplementary Material Section S1) – mean pooling outperforming both attention variants, with plain attention pooling collapsing to near-chance performance on some seeds – runs counter to the broader MIL literature, where learned attention pooling [20] typically outperforms fixed pooling on real, heterogeneous slide-level bags. We attribute this to a mismatch between our synthetic bag construction (Section 3.7) and the setting attention pooling is designed for: our bags are randomly assembled from an already strongly per-instance-informative feature space with a simple majority-vote bag label, so there is little genuine intra-bag heterogeneity for an attention mechanism to exploit, while its additional learned parameters have more opportunity to overfit on comparatively few training bags. We report this result rather than omitting or reframing it, and return to it in Section 7 as a limitation of the current bag construction rather than a claim about attention-based MIL in general.

Grad-CAM/Grad-CAM++ faithfulness (insertion AUC 0.749, deletion AUC 0.316, average confidence drop 55.6%, Section 3.9) is consistent with, rather than exceptional relative to, faithfulness numbers reported elsewhere in the pathology explainability literature; we present it as standard-quality, quantitatively validated saliency rather than as a novel explainability contribution in its own right. The novelty we claim is structural: combining patch-level Grad-CAM saliency with bag-level MIL attention into a two-level explainability output, and reporting faithfulness metrics rather than

presenting heatmaps qualitatively only, which reviews of the field note is uncommon practice (Section 2).

6.4 Unexpected Findings

Four results ran against our prior expectations. First, the strongest in-domain CNN generalized *worst* under our primary (full-image-resize) BACH protocol: EfficientNet-B0 has the highest CNN accuracy and above-average in-domain robustness, yet is the only backbone whose BACH AUC falls below chance and whose decision function collapses to a constant output under that protocol (Section 5.7) – plausibly because its aggressive compound scaling fits the source domain with little redundancy to spare under shift, though our scale-aware tiling ablation (Section 5.7) shows this collapse is substantially protocol-dependent rather than a stable property of the architecture, which tempers this explanation. Second, discrimination and calibration *diverged* under domain shift: ViT-B/16 keeps the best AUC on the PathMNIST test split (0.907) but the *worst* Brier score (0.192), while DenseNet121 calibrates best externally (Section 5.8), so in-domain calibration superiority does not transfer and should be re-checked after any shift. Third, attention-based MIL underperformed mean pooling (Section 6.3), the result most at odds with prior literature [20]. Fourth, ViT-B/16 – the backbone with the best discrimination, calibration, and robustness of the five models – had the *least* faithful explainability output: attention rollout scored lower on every insertion/deletion faithfulness metric than Grad-CAM on EfficientNet-B0 under the identical protocol (Section 5), so a reader should not assume the architecture that wins on discrimination, calibration, and robustness also produces the most trustworthy saliency map; those are separate properties, and by this measure the least robust CNN’s explanations were more faithful than the most robust transformer’s.

7 Limitations and Future Work

We report the following limitations directly rather than minimizing them.

Training-set size. All main-experiment models are trained on 20,000 PCam patches (8% of the full 262K-patch training set), chosen to keep training reproducible on a single-GPU Google Colab environment. A separate, 3-seed learning-curve run (Section 5) – distinct from the 5-seed main experiment and therefore not expected to reproduce its exact numbers – shows ViT-B/16 test AUC saturating between 10,000 and 20,000 patches (0.9587 at 10K, 0.9557 at 20K under that 3-seed run, versus 0.9634 ± 0.0076 for the primary 5-seed result at 20K), which suggests – but does not prove – that training on the full dataset would yield a modest further improvement rather than a qualitatively different result; we did not run the full-dataset experiment due to compute constraints.

Confounded architecture comparison. ViT-B/16 differs from the four CNN backbones not only in architecture but in parameter count (85.8M vs. 4.0–23.5M) and in the learning rate used during fine-tuning (3×10^{-5} vs. 10^{-4} for the CNNs, Section 3.3), and only one transformer architecture is evaluated. We therefore report an architecture-scale *association* under a fixed, documented training protocol, not a fully isolated

causal effect of architecture alone; a controlled ablation matching learning-rate schedule and parameter budget across backbones, and testing at least one additional transformer variant, would be needed to attribute the robustness and calibration gap to architecture with the same confidence with which we report its existence under the tested protocol.

Simulated clinical metadata. PatchCamelyon contains no clinical annotations, so the structured features used in the multimodal fusion demonstration (Supplementary Material Section S1) are simulated, label-conditioned constructs rather than real patient data. This is standard practice for demonstrating multimodal fusion architectures in the absence of paired clinical datasets, following precedent in the computational pathology literature [27, 28], but it means the fusion results demonstrate architectural behavior, not a clinically validated multimodal biomarker.

MIL attention instability. As discussed in Section 6.3, attention-based MIL pooling underperformed and showed high seed-to-seed variance on our synthetic 20-patch bags, in a direction that runs counter to the wider MIL literature. We consider this a property of our bag construction rather than a general critique of attention pooling, but it means our bag-level attention-based explainability output (Section 3.9) should be interpreted with the same caveat.

External validation coverage. Zero-shot external validation (Table 6) covers three datasets: BACH (breast, cross-task), the PathMNIST test split (colorectal, cross-task, derived from CRC-VAL-HE-7K), and CAMELYON17-WILDS (breast lymph node, same-task, cross-hospital), all public patch/image-level research benchmarks. We use *no* hospital-sourced data outside these public benchmarks, perform *no* prospective evaluation, and only CAMELYON17-WILDS provides genuine multi-center same-task validation; BACH and PathMNIST provide none. MHIST (colorectal polyps) was also attempted but could not be retrieved from any queried public mirror and is not included. With only three external datasets and only one same-task cross-hospital comparison, we cannot rule out that any individual result is dataset-specific rather than representative of external generalization in general.

Possible dataset bias and no federated validation. PCam, BACH, PathMNIST (CRC-VAL-HE-7K test split), and CAMELYON17-WILDS are each drawn from a small number of source institutions and scanning setups, and none discloses detailed patient demographic composition alongside the image data; we cannot rule out that some portion of the accuracy, robustness, or calibration differences we report reflect dataset-specific acquisition characteristics or an unmeasured demographic skew rather than a property of the model or architecture. All 25 main-experiment models are trained and evaluated on centrally hosted datasets; we do not evaluate under a federated-learning setting or with data drawn from multiple independent hospital sites simultaneously.

ViT explainability method and HL7 FHIR scope. We evaluate ViT-B/16 with attention rollout (Section 3.9) rather than a gradient-informed method, since gradient-based class-activation mapping does not directly apply to ViT-B/16’s token-based representation; as reported in Section 5, this rollout saliency is measurably less faithful by the insertion/deletion protocol than the CNN’s Grad-CAM, and gradient-informed transformer relevance propagation [29] remains future work rather than something this

revision implements. Separately, the HL7 FHIR R4 export pipeline (Supplementary Material Section S2) has been validated structurally (required field and `resourceType` presence) but not against a live FHIR server or hospital EHR system; real-world interoperability validation is future work.

Patch-clustering assumption in significance tests. The DeLong test, McNemar’s test, and the patch-level bootstrap confidence intervals (Section 3.10) all treat individual test patches as independent observational units. PCam patches can originate from a shared source whole-slide image, and we do not have access to source-slide identifiers for the public PCam distribution used here; if the fixed 8,000-patch test set contains multiple patches from the same slide, these intervals and p -values may be somewhat optimistic relative to a slide-clustered analysis (e.g., a cluster bootstrap by whole-slide image).

No prospective clinical validation. All results in this paper are retrospective, computed on public benchmark datasets; no component of this framework has been evaluated prospectively or in an active clinical workflow.

Corruption-realization matching across backbones, and residual scope of the variance check. Our evaluation code restarts each perturbation protocol’s random state fresh (same fixed seed) at the start of each backbone’s evaluation at a given σ , for both the RGB channel-gain and HED optical-density perturbations (Section 3.4), so the five per-training-seed evaluation passes for that backbone draw corruption realizations in a fixed sequence – and, because every backbone restarts from the same random state in the same sequence, the k -th training-seed’s evaluation pass sees matched corruption realizations across all five backbones at a given σ , which is what makes the paired, seed-matched $\Delta D_{m,\sigma}$ analysis (Section 3.4) a fair comparison. We separately checked, with 10 independently-seeded fixed corruption replicates at $\sigma = 0.20$ (Section 3.4), that corruption-instance variance is negligible relative to training-seed variance for every backbone; we did not repeat this check at $\sigma = 0.05$ – 0.15 or for the HED perturbation, though we have no reason to expect the same RGB-vs-pixel-color-draw mechanism to behave qualitatively differently there.

Future work. Direct extensions include: (i) training on the full 262K-patch Patch-Camelyon training set to test whether the 20K-patch plateau holds at full scale; (ii) whole-slide-level, rather than patch-level, evaluation on CAMELYON17 – this paper evaluates the WILDS benchmark’s pre-extracted patch release (Section 3.12), not full whole-slide images, so slide-level aggregation, true multi-center scanner variation beyond what the patch release captures, and additional single-source datasets remain future extensions; and applying the scale-aware BACH tiling protocol (Section 3.12) with overlapping tiles and a learned (rather than mean/max) aggregator, closer to genuine MIL over real tissue than either the full-image-resize protocol or the fixed non-overlapping tiling reported here; (iii) replacing simulated clinical features with real, paired clinical metadata through a hospital or biobank partnership; (iv) re-evaluating attention-based MIL pooling on larger, more heterogeneous slide-level bags constructed from genuine whole-slide data; (v) implementing gradient-informed transformer relevance propagation [29] for ViT-B/16, motivated directly by the faithfulness gap we measured between attention rollout and CNN Grad-CAM under the

identical insertion/deletion protocol (Section 5); (vi) a capacity- and optimization-matched ablation sequence isolating architecture from pretraining data scale and from confounding factors: a common learning-rate validation grid across architectures, a smaller ViT/DeiT variant and a larger CNN closer to ViT-B/16’s parameter count, at least one additional transformer family, an augmentation ablation (no color jitter vs. current color jitter vs. explicit stain-space augmentation), and a frozen-encoder/linear-probe comparison to separate representation robustness from fine-tuning optimization effects; (vii) a controlled normalization ablation (batch normalization vs. layer normalization, holding architecture otherwise fixed) to test the normalization hypothesis offered for the BACH collapse (Section 5.7); (viii) a Macenko stain-vector perturbation protocol as a further complementary robustness test alongside the RGB channel-gain and HED optical-density perturbations already evaluated here (Section 3.4), together with brightness/gamma and Gaussian-blur perturbations as secondary acquisition-artifact proxies; (ix) evaluating calibration (ECE, Brier, reliability curves), not only discrimination, as a function of corruption severity, and testing temperature scaling fit on PCam validation data and transferred (without re-fitting on external labels) to BACH/PathMNIST; and (x) a staged clinical deployment pathway – federated or multi-institutional retrospective validation, then shadow-mode deployment alongside routine review, then a prospective multi-site evaluation with a predefined statistical analysis plan – as the appropriate, risk-graded route from the retrospective benchmark evidence in this paper to real clinical use.

8 Conclusion

This paper compared five backbone architectures for histopathology patch classification under a low-resource regime (20,000 PatchCamelyon patches, five seeds, single-GPU Google Colab). A Vision Transformer with only ImageNet-1K pretraining proved competitive with the four CNN backbones on accuracy (AUC 0.9634 vs. 0.9436–0.9589) while showing far less relative AUC degradation under RGB channel-gain corruption (1.5% vs. 7.7–11.0%) – replicated under a second, stain-specific hematoxylin-eosin perturbation – and better in-domain calibration (ECE 0.044 vs. 0.071–0.094). We supported this with a two-level explainability pipeline, a computational-cost characterization, and zero-shot validation on three external datasets. On the cross-task datasets (BACH, PathMNIST) this surfaced a model- and protocol-specific failure mode in one CNN; on the same-task, cross-hospital CAMELYON17-WILDS dataset, all five backbones generalized comparably with no significant ViT-vs-CNN difference, indicating these advantages are clearest under stain/color perturbation and cross-task transfer, not general domain shift. A gated multimodal fusion module and an HL7 FHIR R4 pipeline were also implemented, detailed in the Supplementary Material.

We do not claim to match billion-parameter pathology foundation models on raw accuracy and report that gap. Instead, our contribution is a quantified trade-off: robustness and calibration as the properties that determine whether a pathology AI model survives a real, multi-site workflow. We also report results that complicate a

purely accuracy-driven narrative – the best in-domain CNN generalizing worst externally, and attention-based MIL underperforming mean pooling – favoring transparency over a cleaner but incomplete story. We hope this framework and evaluation protocol are useful to research groups operating under similar computational constraints.

Declarations

Ethics approval and consent to participate

Not applicable. This study is a secondary analysis of publicly available, de-identified benchmark datasets (PatchCamelyon, BACH, PathMNIST test split derived from CRC-VAL-HE-7K, CAMELYON17-WILDS; Sections 3.2, 3.12). No new human participants were recruited and no identifiable patient data were used.

Consent for publication

Not applicable. No individual person’s identifiable data, images, or clinical details are presented in this manuscript.

Availability of data and materials

All datasets used are publicly available: PatchCamelyon via its HuggingFace Datasets mirror ([1aurent/PatchCamelyon](#)); BACH via its HuggingFace Datasets mirror (Section 3.12); PathMNIST via the `medmnist` Python package [25]; and CAMELYON17-WILDS via its HuggingFace Datasets mirror ([jxie/camelyon17_ood_test](#) split). Code and trained model checkpoints are available from the corresponding author on reasonable request and will be made publicly available upon publication.

Competing interests

The authors declare no competing interests.

Funding

The authors received no specific funding for this work.

Authors’ contributions

Abir Hasan: Methodology, software, experimentation, data curation, formal analysis, visualization, and writing-original draft. Sadia Yasmin: Experimentation, validation, data analysis, visualization, and writing. Ripa Akter: Data analysis, visualization, and writing. Md. Abdul Awal: Conceptualization, methodology, supervision, interpretation of results, project administration, and writing-review and editing. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

References

- [1] Niazi MKK, Parwani AV, Gurcan MN (2019) Digital pathology and artificial intelligence. *Lancet Oncol* 20(5):e253–e261. [https://doi.org/10.1016/S1470-2045\(19\)30154-8](https://doi.org/10.1016/S1470-2045(19)30154-8)
- [2] Omar M, Alexanderani MK, Valencia I, Loda M, Marchionni L (2024) Applications of digital pathology in cancer: a comprehensive review. *Annu Rev Cancer Biol* 8(1):245–268. <https://doi.org/10.1146/annurev-cancerbio-062822-010523>
- [3] Veeling BS, Linmans J, Winkens J, Cohen T, Welling M (2018) Rotation equivariant CNNs for digital pathology. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G (eds) *Medical image computing and computer-assisted intervention – MICCAI 2018. Lecture notes in computer science*, vol 11071. Springer, Cham, pp 210–218. https://doi.org/10.1007/978-3-030-00934-2_24
- [4] Bejnordi BE, Veta M, van Diest PJ et al (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318(22):2199–2210. <https://doi.org/10.1001/jama.2017.14585>
- [5] Omoush SA, Alzyoud JAM, El-Omari NKT, Alzyoud AJA (2026) The role of whole slide imaging in AI-based digital pathology: current challenges and future directions – an updated literature review. *J Mol Pathol* 7(1):2. <https://doi.org/10.3390/jmp7010002>
- [6] Rodriguez JPM, Rodriguez R, Silva VWK, Kitamura FC, Corradi GCA, de Marchi ACB, Rieder R (2022) Artificial intelligence as a tool for diagnosis in digital pathology whole slide images: a systematic review. *J Pathol Inform* 13:100138. <https://doi.org/10.1016/j.jpi.2022.100138>
- [7] Dimitriou N, Arandjelović O, Caie PD (2019) Deep learning for whole slide image analysis: an overview. *Front Med (Lausanne)* 6:264. <https://doi.org/10.3389/fmed.2019.00264>
- [8] McGenity C, Clarke EL, Jennings C, Matthews G, Cartlidge C, Freduah-Agyemang H, Stocken DD, Treanor D (2024) Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *NPJ Digit Med* 7:114. <https://doi.org/10.1038/s41746-024-01106-8>
- [9] Wang X, Yang S, Zhang J et al (2022) Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal* 81:102559. <https://doi.org/10.1016/j.media.2022.102559>
- [10] Chen RJ, Ding T, Lu MY et al (2024) Towards a general-purpose foundation model for computational pathology. *Nat Med* 30(3):850–862. <https://doi.org/10.1038/s41591-024-02857-3>

- [11] Xu H, Usuyama N, Bagga J et al (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>
- [12] Song AH, Jaume G, Williamson DFK, Lu MY, Vaidya A, Miller TR, Mahmood F (2023) Artificial intelligence for digital and computational pathology. *Nat Rev Bioeng* 1(12):930–949. <https://doi.org/10.1038/s44222-023-00096-8>
- [13] Tellez D, Litjens G, Bándi P et al (2019) Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal* 58:101544. <https://doi.org/10.1016/j.media.2019.101544>
- [14] Cornish TC, Dayal S, Ferber D, Chiweshe J, Monroe R (2026) The evolving role of DICOM in digital pathology. *J Pathol Inform* 21:100658. <https://doi.org/10.1016/j.jpi.2026.100658>
- [15] Parker CTA, Mendes L, Liu VYT et al (2025) External validation of a digital pathology-based multimodal artificial intelligence-derived prognostic model in patients with advanced prostate cancer starting long-term androgen deprivation therapy: a post-hoc ancillary biomarker study of four phase 3 randomised controlled trials of the STAMPEDE platform protocol. *Lancet Digit Health* 7(7):100885. <https://doi.org/10.1016/j.landig.2025.100885>
- [16] Dosovitskiy A, Beyer L, Kolesnikov A et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR) 2021*. <https://doi.org/10.48550/arXiv.2010.11929>
- [17] Aresta G, Araújo T, Kwok S et al (2019) BACH: grand challenge on breast cancer histology images. *Med Image Anal* 56:122–139. <https://doi.org/10.1016/j.media.2019.05.010>
- [18] Kather JN, Krisam J, Charoentong P et al (2019) Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med* 16(1):e1002730. <https://doi.org/10.1371/journal.pmed.1002730>
- [19] Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. PMLR, pp 1321–1330.
- [20] Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance learning. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR, pp 2127–2136.
- [21] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In:

- Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- [22] Bilal KH, Gelfant K, Manzo A, Reuter V, Hameed M, Hanna MG, Ardon O (2026) Comparative analysis of whole-slide scanner tissue detection algorithms: implications for scan area, scan time, and file size in high-volume digital pathology workflows. *J Pathol Inform* 22:100678. <https://doi.org/10.1016/j.jpi.2026.100678>
 - [23] HL7 International (2019) HL7 FHIR release 4 (R4): fast healthcare interoperability resources. <https://www.hl7.org/fhir/>. Accessed 14 Aug 2026
 - [24] Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp 839–847. <https://doi.org/10.1109/WACV.2018.00097>
 - [25] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, Pfister H, Ni B (2023) MedM-NIST v2: a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci Data* 10:41. <https://doi.org/10.1038/s41597-022-01721-8>
 - [26] Wei J, Suriawinata A, Ren B et al (2021) A petri dish for histopathology image analysis. In: Artificial intelligence in medicine – AIME 2021. Lecture notes in computer science. Springer, Cham, pp 11–24. https://doi.org/10.1007/978-3-030-77211-6_2
 - [27] Chen RJ, Lu MY, Weng WH, Chen TY, Williamson DFK, Manz T, Shady M, Mahmood F (2021) Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 3995–4005. <https://doi.org/10.1109/ICCV48922.2021.00398>
 - [28] Braman N, Gordon JWH, Goossens ET, Willis C, Stumpe MC, Venkataraman J (2021) Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In: Medical image computing and computer-assisted intervention – MICCAI 2021. Lecture notes in computer science, vol 12905. Springer, Cham, pp 667–677. https://doi.org/10.1007/978-3-030-87240-3_64
 - [29] Chefer H, Gur S, Wolf L (2021) Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 782–791. <https://doi.org/10.1109/CVPR46437.2021.00084>
 - [30] Ruifrok AC, Johnston DA (2001) Quantification of histochemical staining by color deconvolution. *Anal Quant Cytol Histol* 23(4):291–299.

- [31] Abnar S, Zuidema W (2020) Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp 4190–4197. <https://doi.org/10.18653/v1/2020.acl-main.385>
- [32] Bandi P, Geessink O, Manson Q et al (2018) From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans Med Imaging 38(2):550–560. <https://doi.org/10.1109/TMI.2018.2867350>
- [33] Koh PW, Sagawa S, Marklund H et al (2021) WILDS: a benchmark of in-the-wild distribution shifts. In: Proceedings of the 38th International Conference on Machine Learning (ICML). PMLR, pp 5637–5664.