MULTIMAGNET: MULTI-MAGNIFICATION FUSION WITH AVAILABILITY MASKING, MAGNITUDE DROPOUT, AND FEATURE ALIGNMENT REGULARISATION FOR ROBUST BREAST CANCER HISTOPATHOLOGY CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated breast cancer diagnosis from haematoxylin and eosin (H&E) histopathology slides is limited by a magnification-scale gap: almost all deep learning systems train at a single, arbitrarily chosen zoom level, discarding the multi-scale information that pathologists routinely combine across $40\times$, $100\times$, $200\times$, and $400\times$ magnification. We introduce MULTIMAGNET, a four-stream late-fusion framework that jointly encodes all four magnifications through frozen ImageNet-pretrained ResNet-50 backbones augmented with three components: (i) *availability masking*, an explicit binary signal distinguishing a genuinely absent stream from an informative zero-valued embedding; (ii) *magnitude dropout*, which stochastically erases entire streams during training ($p = 0.10$), implicitly ensembling over all $2^4 - 1 = 15$ missing-view patterns at essentially zero extra cost; and (iii) *feature alignment regularisation* (FAR), an ℓ_2-alignment auxiliary loss enforcing coherence between co-available stream embeddings. On BreakHis (7,909 images, 82 patients), under strict patient-level partitioning, naive four-stream fusion already exceeds the strongest single-stream baseline ($200\times$, AUC $= 0.890$) by 4.0 AUC points. Adding our proposed components yields a set of trade-offs rather than a single optimal model: magnitude dropout provides the best robustness to missing views, while a gated variant (E10) reaches the highest full-stream AUC-ROC (0.935 ± 0.003) but is fragile under partial availability. Transfer to the BACH dataset yields a macro AUC of 0.882 ± 0.016; because BACH is single-magnification and the task is four-class, this is not directly comparable to the binary BreakHis result, but demonstrates retained discriminative capacity out-of-distribution. We evaluate the framework's clinical reliability using Monte Carlo Dropout uncertainty decomposition, Mahalanobis out-of-distribution detection (AUROC $= 1.000$), temperature calibration, and stream attribution methods, resulting in a robust suite of engineering choices deployable at an inference cost of ≈ 30 ms per sample on a single GPU.

1 INTRODUCTION

Breast cancer is the most commonly diagnosed malignancy worldwide, with roughly 2.3 million new cases and 685,000 deaths in 2020 (Sung et al., 2021). Survival has improved markedly where organised screening exists, but prognosis remains poor in resource-constrained regions such as sub-Saharan Africa, where patients frequently present with advanced-stage disease and pathologists are scarce (Wilson et al., 2018). Histopathological examination of H&E-stained biopsies remains the diagnostic gold standard, with pathologists examining specimens at multiple magnifications, $\mathcal{M} = \{40\times, 100\times, 200\times, 400\times\}$. Low magnification reveals glandular architecture and tumour boundaries, intermediate magnification the nuclear-to-cytoplasmic ratio, and high magnification nuclear atypia and mitotic activity. Each magnification provides partially non-redundant information, and integrating these multiscale features is a defining aspect of expert histopathological interpretation (Spanhol et al., 2016).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Nevertheless, most deep learning systems for breast histopathology are trained and evaluated at a single, arbitrarily chosen magnification (Spanhol et al., 2016; Bardou et al., 2018). This *magnification-scale gap* has two consequences. First, no single magnification captures the full diagnostic picture, which imposes a ceiling on achievable accuracy. Second, scanner faults, protocol variations, or incomplete slide preparation may render one or more magnifications unavailable for a given specimen; therefore, a deployable system should degrade gracefully rather than fail when such information is missing.

We address both problems with MULTIMAGNET, a four-stream late-fusion framework built for modest hardware that fuses spatially unregistered morphological features across scales. A core contribution is a comprehensive 15-pattern missing-view evaluation protocol, which, unlike the complete- or single-dropout evaluations common in multimodal work, exposes a sharp robustness gap between fusion variants (Sec. 5.3). Our contributions are:

1. **Architecture.** A four-stream MAGENCODER design (frozen ImageNet ResNet-50 backbones with trainable projection heads) fused either by masked concatenation or by masked soft-attention gating (AUC-ROC 0.935 on BreakHis; Sec. 3).
2. **Availability masking and magnitude dropout.** An explicit presence signal that separates a missing stream from a zero-valued embedding, combined with stream-level stochastic erasure ($p = 0.10$) that implicitly ensembles over all $2^4 - 1 = 15$ non-empty stream subsets at no inference cost.
3. **Feature alignment regularisation (FAR).** An ℓ_2 alignment loss over co-available stream embeddings to enforce geometric consistency.
4. **Evaluation under patient-level splitting.** A controlled ablation (E1–E10) against fully trainable CSPResNet50 baselines exposing clear trade-offs: the gated model achieves the highest full-stream AUC but is the most fragile under partial availability, and temperature scaling does not repair its calibration.

2 RELATED WORK

Histopathology classification and magnification. Spanhol et al. (2016) introduced BreakHis and showed that accuracy depends on magnification. Bayramoglu et al. (2016) trained one magnification-independent CNN across scales, and Gupta & Bhavsar (2017) asked whether magnification matters at all; both still process one magnification at a time. CNN transfer from ImageNet is now standard in this low-data regime (Bardou et al., 2018; Litjens et al., 2017; He et al., 2016). A recurring methodological problem is *image-level* splitting: BreakHis contains many images per patient, so a random image-level split leaks patient identity across partitions and inflates accuracy. We use patient-level partitions throughout (see Appendix A).

Multi-scale and attention-based aggregation. Whole-slide models such as CLAM, DSMIL and HIPT aggregate patches or scales with attention or hierarchical transformers (Lu et al., 2021; Li et al., 2021; Chen et al., 2022), and attention-based MIL (Ilse et al., 2018) aggregates instances within one resolution; these are heavy for resource-limited deployment and assume every scale is present. Our gate adapts attention-based pooling to attend over magnification *streams* and is close in spirit to gated multimodal units (Arevalo et al., 2017); we adopt late fusion, which is a common strategy for views with dissimilar statistics (Baltrušaitis et al., 2019).

Missing views and modality robustness. Wang et al. (2020b) showed that multi-stream networks overfit unevenly across streams, and Ma et al. (2022) showed that multimodal models are sensitive to missing modalities. ModDrop (Neverova et al., 2016), an application of dropout (Srivastava et al., 2014) at the modality level, is the closest antecedent of our magnitude dropout; to our knowledge a systematic 15-pattern missing-view sweep has not been reported for BreakHis.

Reliability. We combine MC-Dropout (Gal & Ghahramani, 2016) with the aleatoric/epistemic decomposition of Kendall & Gal (2017), temperature scaling (Guo et al., 2017), Mahalanobis OOD detection (Lee et al., 2018), and Grad-CAM, Integrated Gradients and SHAP (Selvaraju et al., 2017; Sundararajan et al., 2017; Lundberg & Lee, 2017) (definitions in Appendix D). Pathology foundation models (Chen et al., 2024; Lu et al., 2024) are complementary to our ImageNet backbones (Sec. 6).

3 METHOD

3.1 ARCHITECTURE

MAGENCODER. A magnification encoder is $E_k = \psi_k \circ \phi$, where $\phi : \mathbb{R}^{3\times224\times224} \to \mathbb{R}^{2048}$ is a frozen ImageNet-pretrained ResNet-50 (He et al., 2016) with global average pooling (details in Appendix B) and ψ_k is a stream-specific trainable projection head,

$$\psi_k(I) = \mathrm{ReLU}\Big(\mathrm{BN}\big(W_k^{(2)}\,\mathrm{Drop}_{0.2}\big(\mathrm{ReLU}(\mathrm{BN}(W_k^{(1)}\phi(I) + b_k^{(1)}))\big) + b_k^{(2)}\big)\Big) \in \mathbb{R}^d, \qquad d = 128, \tag{1}$$

with $W_k^{(1)} \in \mathbb{R}^{512\times2048}$ and $W_k^{(2)} \in \mathbb{R}^{d\times512}$. Each head has 1,116,032 parameters (four heads: $\approx 4.46 \times 10^6$) against $\approx 23.5 \times 10^6$ frozen backbone weights per stream. Streams share architecture and pretraining but not projection weights, so each learns a magnification-specific subspace on top of shared ImageNet features.

Fusion with availability masking. MULTIMAGNET maps four images $(I^1, \ldots, I^4)$ and an availability mask $\mathbf{m} \in \{0, 1\}^4$ to class logits $\mathbf{z} \in \mathbb{R}^2$. Each embedding $f^k = E_k(I^k)$ is masked and the mask itself is appended,

$$\mathbf{g} = [\,m_1 f^1;\, m_2 f^2;\, m_3 f^3;\, m_4 f^4;\, \mathbf{m}\,] \in \mathbb{R}^{4d+4}, \tag{2}$$

which a three-layer MLP ($516 \to 256 \to 128 \to 2$; BatchNorm, ReLU, dropout 0.3) maps to $\mathbf{z}$; the malignancy probability is $\hat{y} = \mathrm{softmax}(\mathbf{z})_1$. Appending $\mathbf{m}$, rather than leaving the classifier to infer absence from an all-zero block, is what lets it distinguish a genuinely missing stream from one whose embedding is near zero: this is the *availability masking* component evaluated in Sec. 5.3.

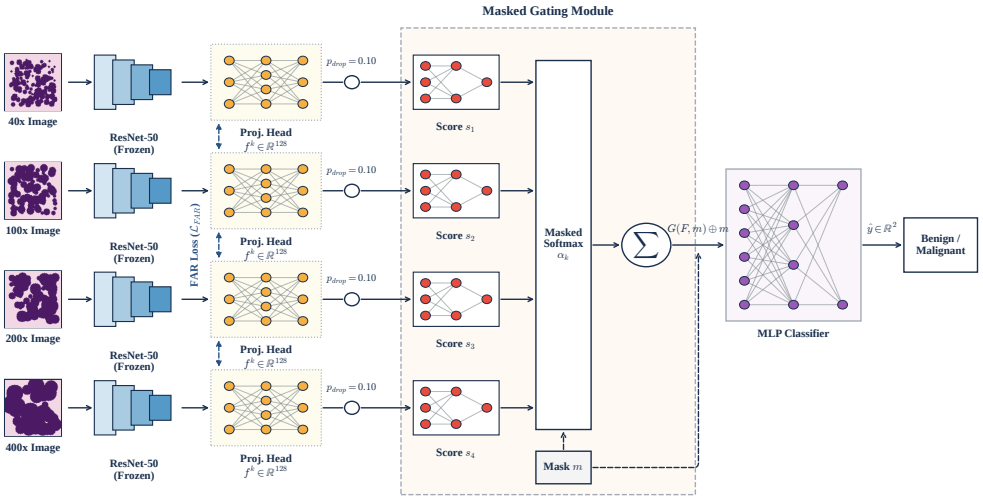


Figure 1: MULTIMAGNET in its MULTIMAGNETGATED (E10) configuration. Each of four magnification images enters a MAGENCODER: a frozen ResNet-50 (locked) followed by a trainable projection head producing $f^k \in \mathbb{R}^{128}$. During training, magnitude dropout ($p = 0.10$) stochastically zeroes whole streams and FAR (dashed double arrows) pulls co-available embeddings together. Per-stream scores s_k feed a masked softmax α_k (unavailable streams receive $-\infty$), and the convex combination $\sum_k \alpha_k f^k$, concatenated with $\mathbf{m}$, is classified by an MLP. Ablations E1–E9 replace the shaded gating block by the flat concatenation of Eq. (2).

Masked gating (MULTIMAGNETGATED, E10). With $\mathbf{F} = [f^1; \ldots; f^4] \in \mathbb{R}^{4\times d}$, the gate scores each stream with a small MLP and normalises over available streams only,

$$s_k = W_g^{(2)}\mathrm{ReLU}(W_g^{(1)} f^k + b_g^{(1)}) + b_g^{(2)}, \quad \alpha_k = \frac{m_k \exp(s_k)}{\sum_{j=1}^4 m_j \exp(s_j)}, \quad G(\mathbf{F}, \mathbf{m}) = \sum_{k=1}^4 \alpha_k f^k, \tag{3}$$

with $W_g^{(1)} \in \mathbb{R}^{64\times d}$, $W_g^{(2)} \in \mathbb{R}^{1\times64}$; $\alpha_k = 0$ exactly for absent streams (equivalent to setting $s_k = -\infty$). The classifier receives $[G(\mathbf{F}, \mathbf{m}); \mathbf{m}] \in \mathbb{R}^{d+4}$. Note that when a single stream is active, α is one-hot and G reduces to that stream's embedding, a property that matters for robustness (Sec. 5.3).

3.2 Objectives and Training

Classification loss. We use label-smoothed cross-entropy (Szegedy et al., 2016; Müller et al., 2019) with $\varepsilon = 0.05$ and correct class imbalance with effective-number-of-samples weighted sampling (Cui et al., 2019). Writing $q_{i,c} \in \{0, 1\}$ for the one-hot label of specimen i, $p_{i,c} = \mathrm{softmax}(\mathbf{z}_i)_c$ for the predicted probability of class $c \in \{0, 1\}$, $C = 2$ for the number of classes, and w_c for the effective-number-of-samples class weight, the loss over a mini-batch $\mathcal{B}$ is

$$\mathcal{L}_{\mathrm{LSCE}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{c=0}^{1} w_c \Big(q_{i,c}(1 - \varepsilon) + \frac{\varepsilon}{C} \Big) \log p_{i,c}. \tag{4}$$

FAR. FAR penalises disagreement between the embeddings of co-available magnifications of the same specimen. Let $\hat{f}_i^k = f_i^k / |f_i^k|_2$ and $\mathcal{C}_i = \{(j, k) : j < k,\ m_{i,j} = m_{i,k} = 1\}$; over the same mini-batch,

$$\mathcal{L}_{\mathrm{FAR}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{|\mathcal{C}_i|} \sum_{(j,k) \in \mathcal{C}_i} |\hat{f}_i^j - \hat{f}_i^k|_2^2. \tag{5}$$

Since $|\hat{f}^j - \hat{f}^k|_2^2 = 2(1 - \cos\theta_{jk})$, minimising $\mathcal{L}_{\mathrm{FAR}}$ maximises the mean pairwise cosine similarity. The total objective is

$$\mathcal{L}_{\mathrm{Total}} = \underbrace{-\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{c=0}^{1} w_c \Big(q_{i,c}(1 - \varepsilon) + \frac{\varepsilon}{C} \Big) \log p_{i,c}}_{\mathcal{L}_{\mathrm{LSCE}}} + \underbrace{\frac{\lambda}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{|\mathcal{C}_i|} \sum_{(j,k) \in \mathcal{C}_i} |\hat{f}_i^j - \hat{f}_i^k|_2^2}_{\lambda\, \mathcal{L}_{\mathrm{FAR}}}, \tag{6}$$

with $\lambda = 0.30$ for E9–E10 and 0 otherwise.

Because Eq. (6) is additive, back-propagation splits linearly into a classification force and an alignment force whose relative size is set by λ. The class weights w_c rescale the error signal inside the logit gradient, and FAR's gradient is orthogonal to $\hat{f}_i^j$: it can only *rotate* an embedding toward its co-available partner, never rescale it, and it vanishes as the streams align, so FAR is self-limiting rather than collapsing streams to a point (derivation in Appendix C). This matches the small, saturating AUC-ROC shift ($0.928 \to 0.926$) when FAR is added to masking and dropout (E8→E9, Table 1).

Proposition 1 (FAR geometric bound)**.** *Draw $i \in \mathcal{B}$ uniformly and then a pair $(j, k) \in \mathcal{C}_i$ uniformly, and let $c = \cos\theta_{jk}$ under this draw. If $\mathcal{L}_{\mathrm{FAR}}(\mathcal{B}) \le \delta$, then (i) $\mathbb{E}[c] = 1 - \mathcal{L}_{\mathrm{FAR}}/2 \ge 1 - \delta/2$; (ii) for every $t \in (0, 2]$, $\Pr[c \le 1 - t] \le \delta/(2t)$; and (iii) for $\delta \le 2$, since the embeddings are ReLU outputs (so $c \in [0, 1]$), $\mathbb{E}[\theta_{jk}] \le \arccos(1 - \delta/2)$.*

Proof. (i) uses the identity $|\hat{f}^j - \hat{f}^k|_2^2 = 2(1 - c)$. (ii) is Markov's inequality applied to the non-negative variable $1 - c$. (iii) applies Jensen's inequality to the concave function $\arccos$ on $[0, 1]$. □

Hence FAR bounds the mean angular deviation among co-available streams, which stabilises downstream gating; for example, $\delta = 0.10$ ensures a mean angular deviation $\le 18.2°$.

Magnitude dropout. During training, stream k of example i is retained independently with probability $1 - p_{\mathrm{drop}}$,

$$m_{i,k}^{\mathrm{drop}} \sim \mathrm{Bernoulli}(1 - p_{\mathrm{drop}}), \qquad p_{\mathrm{drop}} = 0.10, \qquad m_{i,k}^{\mathrm{eff}} = m_{i,k}\, m_{i,k}^{\mathrm{drop}}, \tag{7}$$

with rejection sampling so that at least one stream stays active. The induced number of active streams is $P(4) = 0.656$, $P(3) = 0.292$, $P(2) = 0.049$, $P(1) = 0.004$. As in ModDrop (Neverova et al., 2016), this implicitly ensembles the network over all $2^4 - 1 = 15$ non-empty subsets with no extra parameters or inference cost.

Optimisation. All models use AdamW (Loshchilov & Hutter, 2019) (learning rate 10^{-4}, weight decay 10^{-4}, gradient-norm clipping at 1.0) with cosine annealing to $0.05\times$ the initial rate over 20 epochs and no early stopping; the checkpoint with the best validation AUC-ROC is evaluated on the test partition. Full hyperparameters are in Appendix Table 3.

4 EXPERIMENTAL SETUP

Data and splits. BreakHis (Spanhol et al., 2016) contains 7,909 H&E images (700×460 px; 2,480 benign, 5,429 malignant) from 82 patients with a binary label and one of eight subtype labels per patient. All partitions are made at the *patient* level: no patient contributes images to more than one of the train, validation and test partitions (checked by an assertion in our pipeline), and the held-out test partition contains 1,288 examples (318 benign, 970 malignant). Images are resized to 224×224 and augmented independently per stream (Appendix A). BACH (Aresta et al., 2019) contains 400 images (100 each of Normal, Benign, In Situ, Invasive) acquired at a single magnification. For transfer, only the $200\times$ stream is active ($\mathbf{m} = [0, 0, 1, 0]$), the ResNet-50 backbones stay frozen, and the trainable modules on top (projection heads, gate and classifier; $\approx 4.5 \times 10^6$ parameters) are fine-tuned for 20 epochs at learning rate 5×10^{-5} on a BACH split of 272/48/80 train/validation/test images.

Configurations. Table 1 lists all models. CSPResNet50 (Wang et al., 2020a) is trained end-to-end (no frozen weights, $\approx 23.9 \times 10^6$ parameters) at a single magnification; it tests whether the gains come from multi-scale fusion or merely from a more flexible single-scale network. E1–E4 are single-stream MULTIMAGNET models with a frozen backbone; E5–E10 add one component at a time to naive four-stream concatenation.

Metrics and protocols. AUC-ROC is the primary metric; we also report accuracy, macro-F1, malignant-class sensitivity, the Matthews correlation coefficient (MCC, robust to class imbalance; Chicco & Jurman, 2020) and 15-bin ECE. Unless stated otherwise, numbers are means $\pm$ s.d. over three seeds $\{41, 42, 44\}$ on the test partition. *Cross-validation* (E10 only) uses five class-stratified patient-level folds (15 fold-seed runs; Appendix A.1); the runs share folds, so the reported bootstrap CI understates uncertainty from patient sampling. *Missing-view sweep:* each of the $2^4 - 1 = 15$ non-empty subsets $\mathcal{S}$ is evaluated at test time by setting $m_k = \mathbf{1}[k \in \mathcal{S}]$ and zeroing the excluded inputs; we report minimum, mean and spread ($\max - \min$) of accuracy over subsets. *Reliability:* temperature scaling and MC-Dropout (50 passes) follow Appendix D, and the Mahalanobis OOD detector is defined in Appendix D.1. All timings use one NVIDIA A100-SXM4-40GB GPU (PyTorch 2.7.1).

5 RESULTS

Table 1: Main results and ablation on the BreakHis test partition (1,288 examples), mean $\pm$ s.d. over three seeds. MD = magnitude dropout; Sens. = malignant-class sensitivity; ECE is the pre-calibration expected calibration error (mean over seeds, lower is better). CSP rows are single-magnification, fully trainable CSPResNet50 models. Bold marks the best value per column.

Config	Description	Mask	MD	FAR	AUC-ROC	Accuracy	Macro-F1	MCC	Sens.	ECE
CSP-40	CSPResNet50, 40×	✗	✗	✗	0.835 ± 0.014	0.738 ± 0.018	0.728 ± 0.016	0.487 ± 0.023	0.699	0.057
CSP-100	CSPResNet50, 100×	✗	✗	✗	0.833 ± 0.017	0.772 ± 0.018	0.733 ± 0.015	0.471 ± 0.033	0.856	0.104
CSP-200	CSPResNet50, 200×	✗	✗	✗	0.814 ± 0.007	0.778 ± 0.012	0.738 ± 0.009	0.488 ± 0.023	0.885	0.076
CSP-400	CSPResNet50, 400×	✗	✗	✗	0.836 ± 0.006	0.769 ± 0.017	0.711 ± 0.033	0.456 ± 0.048	0.914	0.113
E1	40× only	✗	✗	✗	0.838 ± 0.016	0.774 ± 0.032	0.746 ± 0.026	0.501 ± 0.039	0.797	**0.042**
E2	100× only	✗	✗	✗	0.885 ± 0.015	0.819 ± 0.016	0.794 ± 0.017	0.592 ± 0.033	0.839	0.061
E3	200× only	✗	✗	✗	0.890 ± 0.016	0.858 ± 0.029	0.825 ± 0.033	0.656 ± 0.068	0.927	0.095
E4	400× only	✗	✗	✗	0.859 ± 0.010	0.829 ± 0.024	0.789 ± 0.029	0.582 ± 0.058	0.910	0.056
E5	All-4, concatenation	✗	✗	✗	0.930 ± 0.002	0.892 ± 0.005	0.865 ± 0.007	0.738 ± 0.013	0.960	0.051
E6	E5 + MD	✗	✓	✗	0.932 ± 0.007	$\mathbf{0.898 \pm 0.006}$	0.874 ± 0.007	$\mathbf{0.754 \pm 0.016}$	**0.961**	0.060
E7	E5 + mask	✓	✗	✗	0.928 ± 0.005	0.893 ± 0.007	0.867 ± 0.008	0.740 ± 0.016	0.958	0.046
E8	E5 + mask + MD	✓	✓	✗	0.928 ± 0.001	0.895 ± 0.010	0.870 ± 0.014	0.746 ± 0.025	0.958	0.066
E9	E8 + FAR	✓	✓	✓	0.926 ± 0.001	0.897 ± 0.004	$\mathbf{0.875 \pm 0.004}$	0.753 ± 0.007	0.948	0.056
E10	Gated fusion + FAR	✓	✓	✓	$\mathbf{0.935 \pm 0.003}$	0.897 ± 0.009	$\mathbf{0.875 \pm 0.009}$	$\mathbf{0.754 \pm 0.019}$	0.943	0.111

5.1 SINGLE-MAGNIFICATION BASELINES AND FUSION

Single streams. Table 1 shows a non-monotonic ranking over magnifications: $200\times$ (E3) is the strongest single stream (AUC 0.890, MCC 0.656), followed by $100\times$ (0.885), $400\times$ (0.859) and $40\times$ (0.838). This is consistent with pathological practice, since $200\times$ resolves nuclear-to-cytoplasmic ratio while retaining stromal context; $400\times$ resolves finer detail but its texture statistics lie furthest from ImageNet, a plausible but untested explanation for its lower AUC. Every fully trainable CSPResNet50 (AUC 0.814–0.836) is below its frozen-backbone counterpart; at $200\times$, E3 exceeds CSP-200 by 7.6 AUC points and 0.168 MCC. In this constrained-data setting, frozen ImageNet

features refined by a small head outperformed end-to-end training of a comparable-size backbone (which was not exhaustively tuned), suggesting that large-scale pretraining with lightweight fusion is data-efficient for histopathology. Our single-stream accuracies (77.4–85.8%) are in line with the 80–85% of the original BreakHis benchmark (Spanhol et al., 2016), which also used patient-level partitions; higher figures under image-level splitting are not comparable because they mix patients across partitions.

Fusion. The largest single gain is from the best single stream to naive four-stream fusion (E3→E5): +4.0 AUC points and +0.082 MCC. Every one of the 18 fusion runs (E5–E10; per-seed AUC 0.924–0.941) exceeds every E3 run (0.868–0.906). With three seeds a paired test is under-powered (the sign-flip permutation test cannot fall below $p \approx 0.25$, and Bonferroni-adjusted paired t-tests are not significant), but the effect sizes are substantial: Cohen's d for the AUC gain over E3 lies between 1.8 and 2.7 for all six fusion variants. Since CSPResNet50 has more trainable capacity than any single-stream model here and still trails, the improvement is attributable to multi-scale fusion rather than model size. ROC curves (Appendix Figure 7) show the fusion gain concentrating at the moderate false-positive rates relevant to screening.

Ablation. Magnitude dropout alone (E6) gives the highest accuracy and MCC and improves specificity from 0.736 to 0.754, i.e. stream erasure mainly curbs false positives. The availability mask alone (E7) is marginally worse under full-stream evaluation (0.928), as expected: with all streams present it conveys no new information, and its value appears under partial availability (Sec. 5.3). Adding FAR (E9) lowers AUC slightly (0.926) but yields the smallest seed variance in the study (s.d. 0.001 AUC, 0.007 MCC), a bias-variance trade-off: the alignment penalty constrains the shared space and reduces sensitivity to initialisation. The gated model (E10) attains the highest AUC (0.935 ± 0.003). However, the spread among E5–E10 (at most 0.009 AUC) is of the order of the seed s.d., and E8 and E9 differ by less than their seed variability, so we frame our findings as engineering trade-offs between discrimination, robustness and stability (Secs. 5.3–5.4). Clinically, E1 has a single-stream sensitivity of 0.797, missing over 20% of malignancies, an unacceptable false-negative rate on its own. Naive fusion (E5) raises sensitivity to 0.960, while E10 balances a high sensitivity (0.943) with improved specificity (0.755); the MCC rises from 0.501 (E1) to 0.738 (E5) and 0.754 (E10), summarising this dual improvement in one prevalence-insensitive number.

5.2 Cross-validation

Five-fold patient-level cross-validation of E10 gives AUC 0.919 ± 0.064 (95% CI $[0.883, 0.955]$), accuracy 0.857 ± 0.068 and macro-F1 0.833 ± 0.077 (s.d. over 15 fold-seed runs; per-fold values in Appendix Table 2). The fixed-split AUC of 0.935 lies inside this interval, so the single split is not a favourable-partition artefact. The spread is real and comes from the small cohort: fold-mean AUC ranges from 0.995 (fold 3) to 0.833 (fold 4), whose seed-wise AUCs span 0.764–0.924, and the s.d. across the five fold means is 0.059. Cross-validation was run for E10 only, so it supports the absolute estimate but not the ranking between configurations.

5.3 Robustness to missing magnification streams

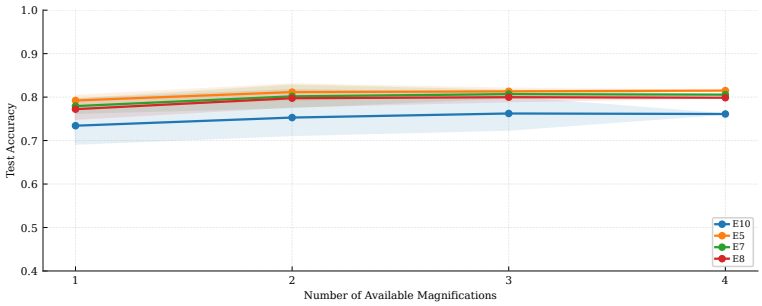


Figure 2: Test accuracy of E5, E7, E8 and E10 versus the number of available magnifications. Lines show the mean over the subsets with n active streams ($n = 1$: 4 subsets; $n = 2$: 6; $n = 3$: 4; $n = 4$: 1) and bands ± 1 s.d. across those subsets.

Figure 2 summarises the 15-pattern sweep. Adding the availability mask (E7 vs. E5) raises the minimum accuracy over subsets from 0.728 to 0.745 by letting the classifier separate an absent

stream from a low-valued one, and adding magnitude dropout (E8) raises it further to 0.759 and the mean to 0.848 (E5: 0.825; E7: 0.833); spread over subsets is 0.159 (E5), 0.153 (E7) and 0.151 (E8). Repeated exposure to incomplete views during training thus transfers directly to partial-availability test conditions and is the most effective robustness intervention. Despite the highest full-stream AUC, E10 is the *most fragile* model under partial availability (minimum accuracy 0.651, mean 0.775, spread 0.187): with one active stream the masked softmax of Eq. (3) is one-hot, so the gate degenerates to a pass-through of that stream without the regularisation that magnitude dropout provides to the concatenation models. Our practical recommendation is E10 when all four magnifications are reliably available and E8 otherwise.

5.4 Calibration and uncertainty

Better discrimination does not imply better calibration. E10 has the highest AUC and also the highest pre-calibration ECE among the MultiMagNet models (0.111; Table 1), whereas the concatenation variants E5–E9 have ECE 0.046–0.066, comparable to the single-stream models (0.042–0.095). Temperature scaling does not fix this (Appendix Figure 8): the validation-fitted $T^* = 1.028 \pm 0.081$ is close to 1 and test ECE marginally increases from 0.111 to 0.115, indicating that validation and test patient groups differ in calibration and that a single scalar is insufficient for the strongest configurations. We therefore report pre-calibration ECE and leave more expressive post-hoc methods to future work.

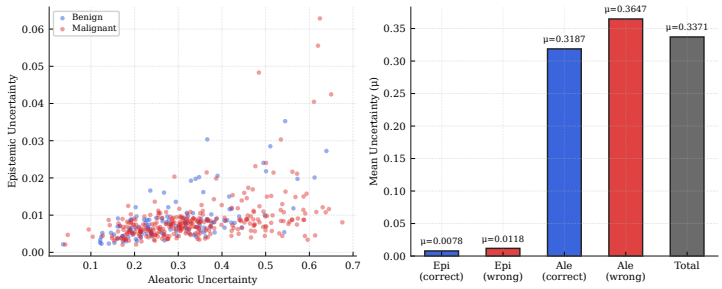
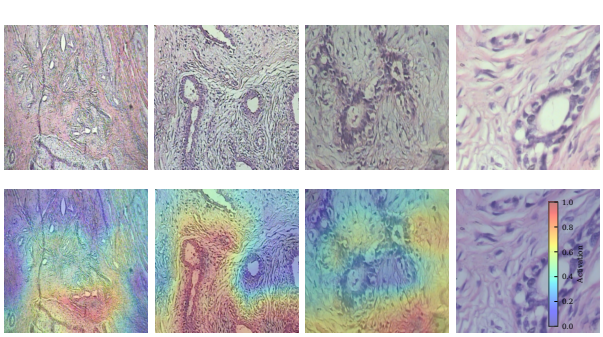


Figure 3: MC-Dropout uncertainty decomposition for E10 (50 passes), split by class and by prediction correctness: epistemic (left) and aleatoric (right) components.

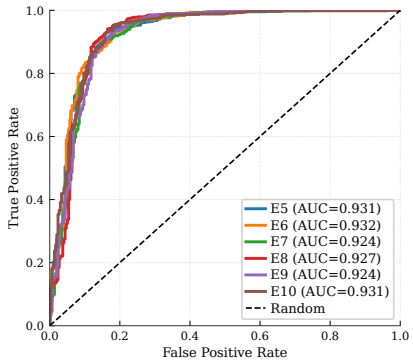
MC-Dropout uncertainty is informative about errors (Figure 3). Misclassified samples carry consistently higher epistemic uncertainty than correctly classified ones, with the clearest separation of the two components, which supports an epistemic threshold for a confidence-gated triage in which high-uncertainty cases are deferred to a pathologist. Malignant samples show higher aleatoric uncertainty than benign ones, consistent with the greater morphological heterogeneity of carcinomas. The distributions still overlap, so uncertainty cannot flag every error and any operating threshold would need prospective calibration on the target population (see Appendix Figure 9 for further failure analysis). For instance, deferring cases where the confidence distance $|p - 0.5| < 0.2$ reduces the autonomous workload by approximately 20.3% while capturing 78.4% of all false negatives.

5.5 Interpretability

Grad-CAM (Figure 4a) shows scale-appropriate saliency, mirroring a pathologist's low-to-high magnification workflow (architecture at $40\times$, epithelial-stromal interface at $100\times$/$200\times$, nuclear pleomorphism and mitotic figures at $400\times$), and Integrated Gradients (Appendix Figure 11b) gives sharper, pixel-level attribution. The same fusion models also achieve consistently high ROC-AUC (0.924–0.932; Figure 4b), corroborating this qualitative saliency pattern with quantitative discrimination. At the stream level, SHAP attributions for E9 rank $40\times$ highest (mean $|\text{SHAP}| = 0.00139$), followed by $100\times$ (0.00135), $200\times$ (0.00113) and $400\times$ (0.00093) (Figure 5, left). This ordering differs from the single-stream AUC ordering (E3>E2>E4>E1) because SHAP measures marginal contribution within the fused representation of E9, not standalone stream power. E10's gate (Appendix Figure 11a), in contrast, downweights $400\times$ and concentrates on $200\times$ for both classes; the two attribution methods probe different quantities. A decision curve analysis (Appendix Figure 11c) further shows positive net clinical benefit across thresholds from 0.05 to 0.85; all interpretability evidence is qualitative (Sec. 6).

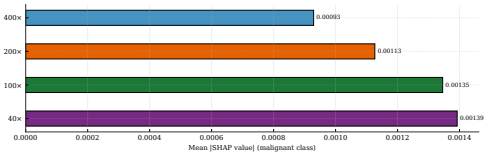


(a) Grad-CAM for a malignant sample, one heatmap per magnification stream. At $40\times$ the encoder attends to irregular tissue architecture, at $100\times/200\times$ to the epithelial-stromal interface, and at $400\times$ to nuclear pleomorphism and mitotic figures, mirroring a pathologist's low-to-high magnification workflow.

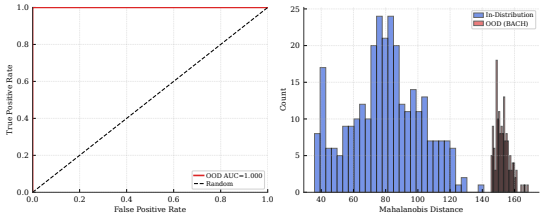


(b) ROC curves for the four-stream fusion models (E5–E10, seed 41), tightly clustered at AUC 0.924–0.932, well above the single-stream envelope (full comparison in Appendix Figure 7).

Figure 4: Qualitative and quantitative evidence for the fused representation. (a) Grad-CAM saliency mirrors a pathologist's low-to-high magnification workflow. (b) The same fusion models attain consistently high discrimination, corroborating the qualitative saliency pattern with quantitative separability.



(a) SHAP stream attribution (E9, malignant class). The marginal ranking $40\times > 100\times > 200\times > 400\times$ differs from the single-stream AUC ordering, indicating that lower magnifications contribute most within the fused representation of E9.



(b) OOD detection via Mahalanobis distance. *Left:* distance histogram for BreakHis test (in-distribution) and BACH (OOD) features. *Right:* ROC curve (AUROC $= 1.000$), confirming reliable separation of the two distributions.

Figure 5: Stream attribution and out-of-distribution detection. SHAP (left) quantifies each stream's marginal contribution in the fused space; the Mahalanobis detector (right) exploits the same feature geometry to flag BACH slides as out-of-distribution with AUROC 1.000.

5.6 Transfer to BACH and out-of-distribution detection

Transferring the BreakHis-trained E10 to four-class BACH (setup in Sec. 4) gives macro one-vs-rest AUC 0.882 ± 0.016 and accuracy 0.625 ± 0.057 on the 80-image test split (learning curves in Appendix Figure 10). Since BACH has a single magnification, only the $200\times$ stream contributes, so E3 on BreakHis (0.890) is the natural reference; the 0.008 AUC difference is descriptive, not controlled — the tasks differ (four-class macro-averaged vs. binary AUC) and the test split is small. This result indicates that the encoders' features remain useful at a new institution, scanner and stain, not that domain shift is absent. A Mahalanobis detector (Appendix D.1) fitted on BreakHis training features separates BreakHis test embeddings from BACH embeddings with AUROC 1.000 (Figure 5b, right): the two distributions are perfectly separated in the learned feature space (Figure 5b, left), consistent with the distinct acquisition conditions. Because BACH differs from BreakHis in every acquisition respect, this is an easy OOD problem; stain- or corruption-shifted BreakHis slides would be a harder and more clinically realistic test.

5.7 Computational cost

The frozen backbones dominate the cost: all MultiMagNet configurations have $\approx 98.7 \times 10^6$ parameters, of which only $4.5\text{–}4.6 \times 10^6$ are trainable, and require 32.7 GFLOPs and 30.0–31.3 ms per sample. The gate adds negligible overhead (30.30 ms for E10 vs. 30.21 ms for E5), so the entire 1,288-example test partition is processed in under 40 s on one GPU (Appendix Table 4) (single-magnification CSPResNet50: 4.1 GFLOPs, 6.8 ms, no fusion gain).

6 Discussion and Limitations

Read together, the results support five conclusions.

(1) Fusion is a structural advantage, not a parameter-count effect. A fully trainable CSPResNet50 cannot match four *frozen* streams at any single magnification (Sec. 5.1), so the 4.0-point AUC gain of naive fusion reflects genuine scale complementarity rather than raw model capacity.

(2) Softmax gating is brittle under missing views, and the mechanism is identifiable. E10 attains the best full-stream AUC yet degrades most sharply as streams are removed (Sec. 5.3). As the normalising denominator in Eq. (3) shrinks toward a single surviving term, $\alpha_k \to 1$, collapsing the gate to a one-hot pass-through whose scale and distribution the classifier never encountered during full-stream training. Magnitude dropout (E8) sidesteps this by exposing the classifier to varying-norm stream combinations throughout training, which is why we recommend E8, not E10, whenever complete four-stream acquisition cannot be guaranteed.

(3) Single-seed reporting would have been misleading. The spread among E5–E10 is at most 0.009 AUC, the same order as seed-to-seed variance, so one seed per configuration could rank these six models differently by chance alone (Sec. 5.1).

(4) The reliability machinery is informative, not decorative. Epistemic uncertainty separates correct from incorrect predictions (Sec. 5.4), Grad-CAM tracks the scale-dependent cues pathologists use (Sec. 5.5), and decision-curve analysis shows positive net clinical benefit across thresholds from 0.05 to 0.85 (Appendix Figure 11c).

(5) The learned features generalise beyond the split and the dataset. Five-fold cross-validation brackets the fixed-split AUC (Sec. 5.2), and the frozen encoders retain discriminative power on the four-class BACH dataset (macro AUC 0.882, Sec. 5.6) despite a magnification, staining and institutional shift.

Together these turn a single leaderboard number into an operating envelope: guaranteed complete four-stream acquisition favours the higher-AUC gated model (E10); uncertain or resource-limited acquisition favours the missingness-robust one (E8).

Limitations. Both datasets originate from single institutions, and BreakHis's 82 patients mean a single borderline case can shift summary statistics — visible in the wide cross-validation fold spread (Sec. 5.2) — while the test partition still contains few patients per histological subtype. Multi-site validation, prioritising sub-Saharan African sites where the diagnostic gap is largest (Sec. 1), is the most pressing next step. Three seeds do not support formal significance testing, and cross-validation was run for E10 alone. Calibration remains incomplete for E9 and E10 (Sec. 5.4); isotonic or Dirichlet calibration, evaluated against decision-curve net benefit rather than ECE alone, is a natural extension. Interpretability evidence (Sec. 5.5) is inspected visually, not validated against pathologist annotations. Our central finding that gating fails under partial availability is shown for one masked-softmax architecture on one dataset; its boundaries need testing with alternative fusion mechanisms (non-normalised gating, cross-attention) and modalities beyond histopathology. Finally, an ImageNet prior is a coarse starting point for histology: pathology foundation models (Chen et al., 2024; Lu et al., 2024) are the most direct route to higher accuracy, and patient-conditional diffusion-based imputation of missing streams is a promising complement to masking and dropout.

7 Conclusion

MultiMagNet tackles the magnification-scale gap by fusing four frozen ResNet-50 streams with availability masking, magnitude dropout, feature alignment regularisation and masked soft-attention gating. Under strict patient-level evaluation, naive four-stream fusion outperforms the strongest single-stream baseline by 4.0 AUC points. Rather than a single optimal architecture, the components offer a navigable set of engineering choices: magnitude dropout and masking (E8) degrade gracefully when views are missing, whereas the gated model (E10) achieves the highest full-stream AUC-ROC (0.935 ± 0.003) but degrades sharply under partial availability and is best reserved for deployments with guaranteed complete data. The reliability study shows where predictions are trustworthy and where they are not, and that calibration needs more than a simple scalar temperature.

Reproducibility Statement

The model, losses and training protocol are specified in Sec. 3; the data partitioning, augmentation, hyperparameters (Appendix Table 3), evaluation protocols and the Mahalanobis detector are specified in Sec. 4 and Appendices A–D.1. All results are means over three fixed seeds $\{41, 42, 44\}$ with reported standard deviations. Source code for training, evaluation and figure generation is provided as supplementary material attached to this submission.

Ethics Statement

This work uses only publicly available, de-identified datasets (BreakHis, BACH) and collected no new human-subjects data. Predictions are intended as decision support within a confidence-gated triage workflow (Sec. 5.4), not as autonomous diagnoses; pathologist review remains mandatory given the residual calibration and interpretability limitations (Sec. 6).

Statement on the Use of Large Language Models

ChatGPT was used to proofread and correct the LaTeX source of this manuscript (notation, formatting, and typesetting) prior to submission. It was not used to design the method, run experiments, or generate reported results.

References

Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, et al. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. Classification of breast cancer based on histology images using convolutional neural networks. *IEEE Access*, 6:24680–24693, 2018.

Neslihan Bayramoglu, Juho Kannala, and Janne Heikkilä. Deep learning for magnification independent breast cancer histopathology image classification. In *International Conference on Pattern Recognition (ICPR)*, pp. 2440–2445, 2016.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16144–16155, 2022.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6, 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pp. 1050–1059, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Vibha Gupta and Arnav Bhavsar. Breast cancer histopathological image classification: Is magnification important? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 17–24, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning (ICML)*, pp. 2127–2136, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2021.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, 2022.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2016.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Fabio Alexandre Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A new backbone that can enhance learning capability of CNN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 390–391, 2020a.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, 2020b.

Michael L. Wilson, Kenneth A. Fleming, Modupe A. Kuti, Lai Meng Looi, Nestor Lago, and Kun Ru. Access to pathology and laboratory medicine services: A crucial gap. *The Lancet*, 391(10133): 1927–1938, 2018.

This appendix is organised in five parts. Appendix A gives full data, preprocessing and cross-validation details; Appendix B specifies the backbone architectures, hyperparameters and computational profile; Appendix C derives the gradient properties of the training objective; Appendix D defines the reliability and interpretability quantities used in Sec. 5.4–5.5; and Appendix E presents supporting figures and qualitative analyses not included in the main text.

A Data, Preprocessing, and Evaluation Protocol

BreakHis composition. The dataset comprises 7,909 images (2,480 benign, 5,429 malignant; 31%/69%) from 82 patients with a long-tailed subtype structure: ductal carcinoma alone contributes 3,451 images from 38 patients, whereas lobular carcinoma (626 images, 5 patients) and phyllodes tumor (453 images, 3 patients) are markedly under-represented. Patient identifiers are parsed from the image filenames. This yields 81 distinct identifiers (24 benign, 57 malignant) for the 82 patients, consistent with a single identifier collision; because a collision can only merge two patients and never split one, partitioning on identifiers cannot leak a patient across partitions and is, if anything, conservative. The five cross-validation folds are stratified by class over these identifiers.

Examples. Each example is a four-view tuple of images of the same slide, one per magnification, grouped by slide and acquisition index; images of different magnifications are not spatially registered. When a magnification is missing for a slide, its stream is zero-filled and masked. The test partition contains 1,288 such examples (318 benign, 970 malignant). During training, the image of each stream is drawn at random among the candidate images of that slide.

Preprocessing and augmentation. Images are resized to 224×224 and normalized with ImageNet statistics $\boldsymbol{\mu} = [0.485, 0.456, 0.406]$, $\boldsymbol{\sigma} = [0.229, 0.224, 0.225]$. Training applies independently to each stream: horizontal and vertical flips ($p = 0.5$ each), rotation $\theta \sim \mathcal{U}(-15^\circ, 15^\circ)$, translation $(\delta_x, \delta_y) \sim \mathcal{U}([-0.10, 0.10]^2)$, Gaussian blur $\sigma \sim \mathcal{U}(0.1, 2.0)$ with probability 0.5, and color jitter (brightness, contrast, saturation, hue) drawn from $\mathcal{U}(-0.15, 0.15)$. Validation and testing use resizing and normalization only.

BACH. The 400 images (100 per class) are split at the patient level into 272/48/80 train/validation/test images; the 80-image test split, which is small, explains the wide seed-to-seed variation of accuracy (0.550–0.688).

A.1 Cross-validation protocol and per-fold results

All 81 distinct patient identifiers extracted from the BreakHis filenames (24 benign, 57 malignant; see above) are stratified by class into five folds of about 16 patients each; with three seeds $\{41, 42, 44\}$ this gives 15 fold-seed runs of E10, evaluated on the corresponding held-out fold. Table 2 reports the per-fold means and standard deviations over seeds, together with the overall mean, standard deviation and bootstrap 95% confidence interval ($B = 1{,}000$ resamples) computed over all 15 runs.

Table 2: Five-fold patient-level cross-validation of E10. Fold entries are means $\pm$ s.d. over seeds $41, 42, 44$; the last two rows use all 15 fold-seed runs (bootstrap CI, $B = 1000$).

Fold	AUC-ROC	Accuracy	Macro-F1
1	0.905 ± 0.021	0.818 ± 0.028	0.775 ± 0.035
2	0.926 ± 0.017	0.798 ± 0.086	0.773 ± 0.077
3	0.995 ± 0.004	0.934 ± 0.060	0.926 ± 0.065
4	0.833 ± 0.082	0.834 ± 0.022	0.799 ± 0.014
5	0.936 ± 0.026	0.902 ± 0.011	0.891 ± 0.012
Mean $\pm$ s.d. (15 runs)	0.919 ± 0.064	0.857 ± 0.068	0.833 ± 0.077
95% bootstrap CI	[0.883, 0.955]	[0.819, 0.895]	[0.790, 0.876]

Fold 3 is the easiest (0.995) and fold 4 the hardest and least stable, consistent with the small-cohort effect discussed in Sec. 5.2. Because the three seeds reuse the same folds, the interval reflects training variability and fold composition but not independent resampling of patients, and should be read as optimistic: it does not capture uncertainty from drawing a different set of patients into each fold.

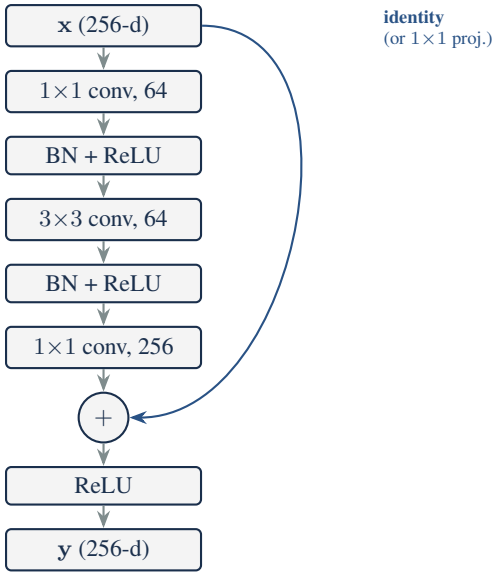
B Model Architecture and Training Configuration

This appendix first specifies the two backbones compared throughout the paper, then reports the full hyperparameter and computational-profile tables that quantify their training and inference cost (Appendix B.1).

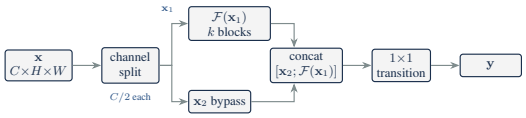
Each stream uses a frozen ImageNet-pretrained ResNet-50 (He et al., 2016): a convolutional stem, four stages of bottleneck blocks (depths 3-4-6-3) and global average pooling, giving $\phi(\mathbf{x}) \in \mathbb{R}^{2048}$. A residual block computes $\mathbf{x}_{l+1} = \sigma(\mathbf{x}_l + \mathcal{F}(\mathbf{x}_l))$ with $\sigma = \mathrm{ReLU}$, and for ResNet-50 the residual branch is the bottleneck (Appendix Figure 6a)

$$\mathcal{F}(\mathbf{x}) = \mathrm{BN}\big(W^{(3)} * \sigma(\mathrm{BN}(W^{(2)} * \sigma(\mathrm{BN}(W^{(1)} * \mathbf{x}))))\big), \tag{8}$$

with $W^{(1)} \in \mathbb{R}^{C/4\times C\times 1\times 1}$, $W^{(2)} \in \mathbb{R}^{C/4\times C/4\times 3\times 3}$, $W^{(3)} \in \mathbb{R}^{C\times C/4\times 1\times 1}$. Batch normalisation (Ioffe & Szegedy, 2015) statistics stay fixed because the backbone is frozen. The CSPResNet50 baseline (Wang et al., 2020a) splits the input of each stage channel-wise into $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{(C/2)\times H\times W}$; only $\mathbf{x}_1$ passes through the residual blocks, $\mathbf{x}_2$ bypasses them, and the outputs are concatenated, $\mathbf{y} = \mathrm{Conv}([\mathbf{x}_2; \mathcal{F}(\mathbf{x}_1)])$ (Appendix Figure 6b). It is trained end to end at a single magnification with the same 3-4-6-3 layout.



(a) Bottleneck residual block ($C = 256$): 1×1 reduction, 3×3 spatial convolution, 1×1 restoration, additive skip connection.



(b) CSP stage: $\mathbf{x}_1$ traverses k bottleneck blocks; $\mathbf{x}_2$ bypasses them and is concatenated at the output before a 1×1 transition.

Figure 6: Core computational units of the two backbones compared in this work. (a) The bottleneck residual block used in every stage of the frozen ResNet-50 MagEncoder backbone. (b) The Cross-Stage-Partial (CSP) stage that replaces each residual stage in the fully-trainable CSPResNet50 baseline. Both backbones share the same macro stem/stage/pooling layout (block counts 3-4-6-3); only the internal stage mechanism differs.

B.1 Hyperparameters and Computational Profile

Table 3 lists every hyperparameter shared across E1–E10; the only settings that vary between configurations are the FAR weight λ and the presence of the gate, masking and magnitude-dropout modules (Sec. 3). Table 4 then breaks down parameter counts, FLOPs and measured latency per configuration, isolating the cost of the frozen backbones from the small trainable head that differs across E1–E10.

C Gradient Analysis of the Training Objective

This appendix derives the two gradient properties summarised at the end of Sec. 3.2. Because Eq. (6) is additive, back-propagation splits linearly, $\nabla_\theta \mathcal{L}_{\mathrm{Total}} = \nabla_\theta \mathcal{L}_{\mathrm{LSCE}} + \lambda \nabla_\theta \mathcal{L}_{\mathrm{FAR}}$, so every SGD

Table 3: Hyperparameters used for all MULTIMAGNET experiments.

Hyperparameter	Value
Image size	224×224
Feature dimension d	128
Projection dropout / classifier dropout	0.20 / 0.30
Magnitude drop probability p_{drop}	0.10
Label smoothing ε	0.05
FAR weight λ	0.30 (E9, E10); 0 (E1–E8)
Optimiser	AdamW, $(\beta_1, \beta_2) = (0.9, 0.999)$
Learning rate / weight decay	10^{-4} / 10^{-4}
LR schedule	cosine annealing to $0.05\,\eta_0$
Gradient clip norm	1.0
Epochs (no early stopping)	20
Cross-validation folds	5
Random seeds	$\{41, 42, 44\}$
Backbone	ResNet-50, ImageNet-pretrained, frozen
Trainable / total parameters (E1–E9)	$\approx 4.63 \times 10^6$ / $\approx 98.7 \times 10^6$
GPU	NVIDIA A100-SXM4-40GB

Table 4: Computational profile of all configurations. Latency is the mean over 50 timed forward passes (20 warm-up passes discarded) on a single GPU for one four-view example.

Config	Description	Total params	Trainable	GFLOPs	Latency (ms)
CSP	CSPResNet50 (one magnification)	23,888,642	23,888,642	4.10	6.80
E1	40× only	98,661,506	4,629,378	32.70	30.20
E2	100× only	98,661,506	4,629,378	32.70	30.02
E3	200× only	98,661,506	4,629,378	32.70	31.34
E4	400× only	98,661,506	4,629,378	32.70	30.16
E5	All-4, concatenation	98,661,506	4,629,378	32.70	30.21
E6	E5 + magnitude dropout	98,661,506	4,629,378	32.70	30.23
E7	E5 + mask	98,662,530	4,630,402	32.70	30.41
E8	E5 + mask + dropout	98,662,530	4,630,402	32.70	30.15
E9	E8 + FAR	98,662,530	4,630,402	32.70	30.32
E10	Gated fusion + FAR	98,530,371	4,498,243	32.70	30.30

step $\theta \leftarrow \theta - \eta\big(\nabla_\theta \mathcal{L}_{\text{LSCE}} + \lambda \nabla_\theta \mathcal{L}_{\text{FAR}}\big)$ sums a classification force and an alignment force whose relative size is set directly by λ.

Classification gradient. Writing $t_{i,c} = q_{i,c}(1-\varepsilon) + \varepsilon/C$, the gradient of Eq. (4) with respect to a logit is

$$\frac{\partial \mathcal{L}_{\text{LSCE}}}{\partial z_{i,d}} = \frac{1}{|\mathcal{B}|}\Big(p_{i,d} \sum_c w_c t_{i,c} - w_d t_{i,d}\Big), \tag{9}$$

where the effective-number weight w_d rescales the error signal *inside* the gradient itself, so a minority-class example exerts a proportionally larger pull on the shared trainable layers at every update, not merely a larger loss value at convergence.

Alignment gradient. Since $\hat{f}_i^k = f_i^k / |f_i^k|_2$, the chain rule through this normalisation gives

$$\frac{\partial}{\partial f_i^j} \big|\hat{f}_i^j - \hat{f}_i^k\big|_2^2 = \frac{2}{|f_i^j|_2}\big(I - \hat{f}_i^j \hat{f}_i^{j\top}\big)\big(\hat{f}_i^j - \hat{f}_i^k\big), \tag{10}$$

which is orthogonal to $\hat{f}_i^j$ by construction. FAR's gradient therefore has no radial component: it can only rotate an embedding across the unit hypersphere toward its co-available partner, never rescale it, which is exactly the angular quantity Proposition 1 bounds. This rotational force also vanishes as $\hat{f}_i^j \to \hat{f}_i^k$, so FAR is self-limiting rather than collapsing streams to a point, consistent with the small, saturating shift in AUC-ROC ($0.928 \to 0.926$) when FAR is added on top of masking and dropout (E8→E9, Table 1).

D Definitions for the Reliability and Interpretability Analyses

This appendix gives the exact formulae behind every reliability and interpretability quantity reported in Sec. 5.4 and Sec. 5.5 (Grad-CAM, Integrated Gradients, SHAP, and the sensitivity/specificity/MCC metrics), followed by a self-contained specification of the Mahalanobis out-of-distribution detector used in Sec. 4.

Temperature scaling and MC-Dropout. A scalar temperature $T^* = \arg\min_{T>0} -\sum_i \log \mathrm{softmax}(\mathbf{z}_i/T)_{y_i}$ is fitted on the validation partition. MC-Dropout uses $M = 50$ stochastic passes with epistemic uncertainty $\mathrm{Var}_t[p_t]$, aleatoric uncertainty $\frac{1}{M}\sum_t p_t(1-p_t)$ and total uncertainty $\bar{p}(1-\bar{p})$, where $\bar{p} = \frac{1}{M}\sum_t p_t$.

Grad-CAM. For stream k, target class c and Stage-4 feature maps $A^{k,m}$, channel importances are $\alpha^c_{k,m} = \frac{1}{H'W'}\sum_{i,j} \frac{\partial S^c}{\partial A^{k,m}_{i,j}}$ and the map is $L^c_k = \mathrm{ReLU}(\sum_m \alpha^c_{k,m} A^{k,m})$, upsampled bilinearly to 224×224 (Selvaraju et al., 2017).

Integrated Gradients. From a black baseline $\mathbf{x}'$, $\mathrm{IG}_i(\mathbf{x}) = (x_i - x'_i)\int_0^1 \partial F(\mathbf{x}' + \alpha(\mathbf{x}-\mathbf{x}'))/\partial x_i \, \mathrm{d}\alpha$, which satisfies completeness, $\sum_i \mathrm{IG}_i(\mathbf{x}) = F(\mathbf{x}) - F(\mathbf{x}')$ (Sundararajan et al., 2017).

SHAP stream attribution. With the four streams as players N, the Shapley value of stream i is $\phi_i = \sum_{S \subseteq N\setminus\{i\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!}[v(S\cup\{i\}) - v(S)]$, where $v(S)$ is the model output using only coalition S (Lundberg & Lee, 2017). Attributions are estimated with a gradient-based explainer over the fused feature vector of E9, so the reported values are means of $|\phi_i|$ for the malignant class.

Metrics. With $\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN}$ defined for the malignant class, $\mathrm{Sens.} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, $\mathrm{Spec.} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$ and $\mathrm{MCC} = \frac{\mathrm{TP\cdot TN - FP\cdot FN}}{\sqrt{(\mathrm{TP+FP})(\mathrm{TP+FN})(\mathrm{TN+FP})(\mathrm{TN+FN})}}$.

D.1 Mahalanobis-distance OOD detection

The detector operates on the 128-dimensional E10 embeddings, needs no retraining, and is fitted once on in-distribution training features. With class means and a tied covariance

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c}\sum_{i:y_i=c} \mathbf{f}_i, \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{N-2}\sum_c \sum_{i:y_i=c} (\mathbf{f}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{f}_i - \hat{\boldsymbol{\mu}}_c)^\top, \tag{11}$$

the per-class distance of a test feature $\mathbf{f}^*$ is $d_c(\mathbf{f}^*) = (\mathbf{f}^* - \hat{\boldsymbol{\mu}}_c)^\top (\hat{\boldsymbol{\Sigma}} + \varepsilon \mathbf{I})^{-1} (\mathbf{f}^* - \hat{\boldsymbol{\mu}}_c)$ with $\varepsilon = 10^{-5}$, and the OOD score is $\delta(\mathbf{f}^*) = \min_{c\in\{0,1\}} d_c(\mathbf{f}^*)$. Samples with $\delta > \tau$ are flagged, with τ set on held-out in-distribution validation data at a 5% false-positive rate. The tied covariance lowers estimation variance relative to per-class covariance estimates, at the cost of presuming a linear class boundary in Mahalanobis space; this assumption is consistent with the clear benign/malignant separation already visible in the learned embedding (main Figure 5, left), though it is not separately validated here. The reported AUROC of 1.000 separates BreakHis test features from BACH features perfectly.

E Additional Results and Qualitative Analysis

This final appendix collects the supporting figures referenced from the main text: full ROC curves, failure-mode and BACH fine-tuning analyses, and the gate weights, Integrated Gradients, and decision-curve panel.

ROC curves (Appendix Figure 7). The fusion curves are tightly clustered, in line with the small AUC differences among E5–E10, and lie above the single-stream curves over most of the operating range. At the lowest false-positive rates the best single streams (E2, E3) are competitive with fusion models, so the fusion gain concentrates at the moderate false-positive rates relevant to a screening threshold.

Calibration (Appendix Figure 8). Temperature scaling leaves the E10 reliability diagram essentially unchanged: the fitted temperature is close to 1 and ECE marginally increases (Sec. 5.4).

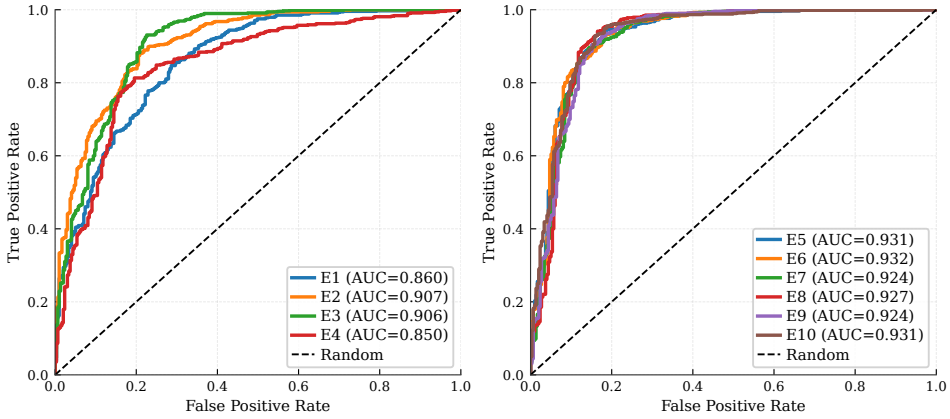


Figure 7: ROC curves on the BreakHis test partition (seed 41). *Left:* single-stream models E1–E4. *Right:* four-stream fusion models E5–E10. Fusion curves are tightly clustered and lie above the single-stream envelope across most operating thresholds; the fusion gain concentrates at the moderate FPR range relevant to screening.

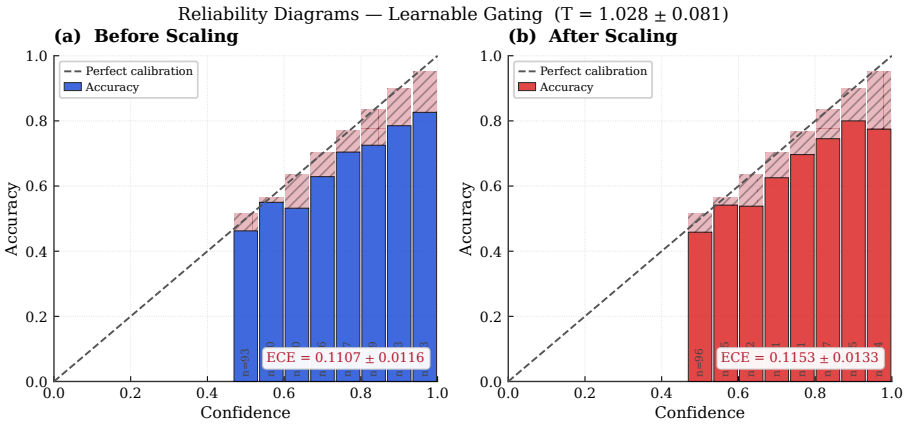


Figure 8: Reliability diagrams for E10 before (a) and after (b) temperature scaling (pooled over three seeds). The fitted temperature is $T^* = 1.028 \pm 0.081$; ECE marginally *increases* from 0.111 to 0.115.

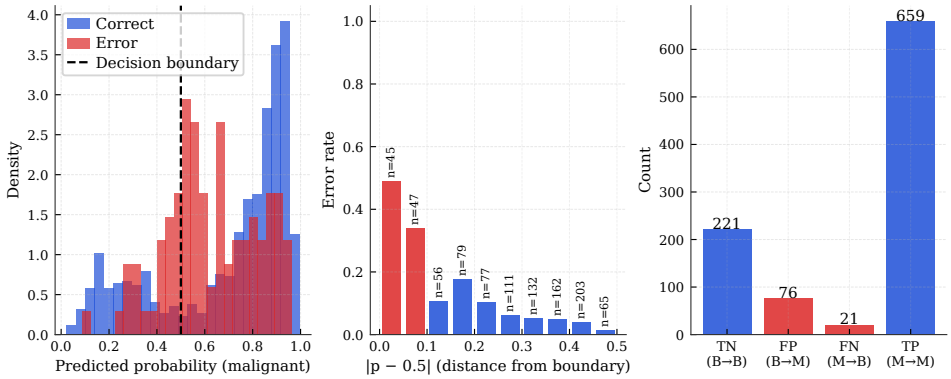


Figure 9: E10 error analysis: distribution of predicted malignancy probability by outcome (left) and error rate as a function of confidence distance $|p - 0.5|$ (right).

Failure analysis (Appendix Figure 9). Errors concentrate in the low-confidence band around the decision boundary, and the error rate falls sharply for $|p - 0.5| > 0.3$. This supports a triage policy that defers low-confidence cases to a pathologist while accepting confident predictions automatically; the fraction that must be deferred at a given target error rate depends on the deployment population and would have to be set prospectively.

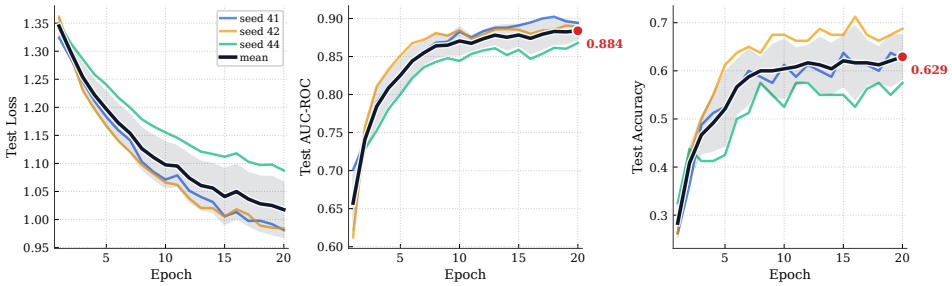
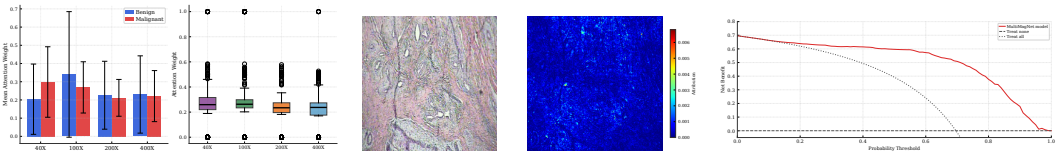


Figure 10: BACH fine-tuning: test AUC per epoch for the three seeds.

BACH fine-tuning (Appendix Figure 10). Test AUC rises quickly during the first epochs and then plateaus at roughly 0.86–0.90 across seeds, so the trainable modules adapt to the new dataset within a few epochs even though the backbones are frozen.



(a) E10 gate weights α_k. (b) Integrated Gradients, $40\times$. (c) Decision curve, E10.

Figure 11: (a) The gate assigns the largest weight to $200\times$ for both classes. (b) Highlighted regions correspond to disrupted glandular organisation and stromal infiltration; attribution is sparser and more spatially precise than Grad-CAM. (c) E10 has positive net clinical benefit across thresholds from 0.05 to 0.85.

Interpretability and clinical utility (Figure 4 and main Figure 5 in the main text; Appendix Figure 11). The SHAP marginal ranking ($40\times > 100\times > 200\times > 400\times$) differs from the single-stream AUC ordering; the gate's preference for $200\times$ reflects a different quantity (attention weight vs. Shapley value), and both are consistent with all four scales contributing to fusion gain. Integrated Gradients satisfies the completeness axiom and corroborates Grad-CAM spatially. The decision curve shows positive net clinical benefit over both default strategies across the plausible threshold range.

OOD and feature space (main Figure 5). The Mahalanobis detector exploits the well-separated class clusters visible in the learned embedding to flag BACH slides perfectly (AUROC 1.000); the linear-boundary assumption this relies on is discussed in Appendix D.1.