Ensemble of Pathology Foundation Models for Fine-Grained Histologic Classification of Glioma

Threya Reddy [1][0009-0007-4541-3019], Arya Pradeep Menon [1][0000-0001-7216-5758], Farzana Rahman [1][0000-0002-4785-5845], Dimitrios Makris [1][0000-0001-6170-0236]

[1]Kingston University, London, UK

K2557789@kingston.ac.uk, A.P.Menon@kingston.ac.uk, Farzana@kingston.ac.uk, D.Makris@kingston.ac.uk

Abstract. Glioblastoma is a highly aggressive primary brain tumor whose diagnosis and grading rely on the identification of heterogeneous histologic sub-regions in digitized tissue sections. The BraTS-Path 2026 challenge frames this as a fine-grained, ten-class classification of H&E-stained image patches. We present a submission built on two complementary pathology foundation models, H-optimus-1 and Virchow2, used as frozen feature extractors. For each backbone we train a lightweight multilayer-perceptron classifier on the extracted embeddings, using focal loss and class-balanced sampling to counter the strong class imbalance of the dataset. We further apply a per-class probability calibration to the H-optimus-1 head and combine its output with that of the Virchow2 head by weighted soft-voting. On the official challenge validation set, the calibrated H-optimus-1 head and the Virchow2 head reach F1 scores of 0.555 and 0.538 respectively, and the weighted ensemble improves over both, attaining an F1 of 0.570, an accuracy of 0.878, an AUC of 0.922, a Matthews correlation coefficient of 0.830, a sensitivity of 0.578, and a specificity of 0.982. The ensemble improves over or matches each individual backbone on every reported metric. Our source code is available at: https://github.com/threyareddy.

Keywords: Computational pathology, Foundational models, Glioma, Digital histopathology, Ensemble learning, Class imbalance, BraTS, H-optimus-1, Virchow2.

1 Introduction

Glioblastoma is the most common malignant primary parenchymal tumor of the brain and carries a grim prognosis, with a median survival of only 12 to 18 months [1]. It is widely infiltrative and pronouncedly heterogeneous in both molecular profile and histopathologic appearance [9], and accurate assessment of that heterogeneity is central to diagnosis, grading and the selection of therapy. In the gold-standard, histopathology-

based approach to tumor diagnosis, expert neuropathologists identify distinct morphopathological features throughout a digitized tissue section [1].

The BraTS-Path 2026 challenge (Task 5) [1] formalizes this problem as a supervised patch-classification task. Annotated regions of H&E-stained, FFPE-digitized tissue sections from the TCGA-GBM and TCGA-LGG collections are divided into equally sized patches, each assigned to one of ten classes: cellular tumor (CT), dense macrophages (DM), infiltration into the cortex (IC), leptomeningeal infiltration (LI), microvascular proliferation (MP), geographic necrosis (NC), lymphocytes (PL), pseudopalisading necrosis (PN), white matter (WM), and none-of-the-above (NOTA). Each patch is treated as an independent case, enabling a fine-grained analysis of the histologic landscape. Fig.1 illustrates the visual diversity and histopathological heterogeneity across tissue sub-regions encountered in glioblastoma specimens.

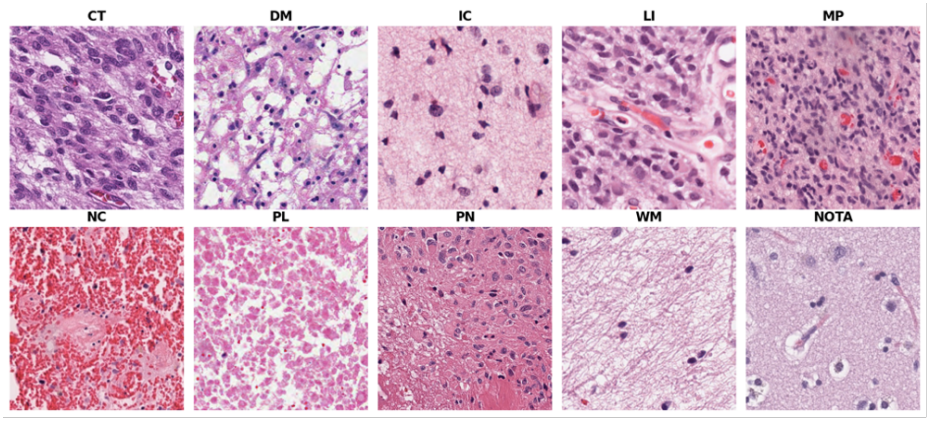

Fig. 1. Typical H&E-stained image patches from the BraTS-Path 2026 dataset, one example patch for each of the ten annotated tissue categories.

Pathology foundation models [3, 10, 15, 21], trained self-supervised on large collections of unlabelled whole-slide images, learn transferable representations capturing both fine-grained cellular morphology and broader tissue architecture. Prior BraTS-Path solutions have used lightweight CNNs such as EfficientNet [16] and MobileNetV2 [4], foundation-model features under long-tailed class distributions [13], contrastive-learning encoders [19], and broader deep-learning pipelines [20]. We employ two such models, H-optimus-1 [2,15] and Virchow2 [21], as frozen feature extractors and train only lightweight classifier heads on their embeddings.

Our contribution is threefold. We train lightweight, class-balanced heads on the embeddings of two frozen pathology foundation models, we introduce a per-class probability calibration on the H-optimus-1 head that improves recognition of rare classes. And, because the encoders are never updated, embeddings are computed once and cached, reducing the cost of training, calibration and ensembling enough to characterise the model across seeds.

2 Methods

Self-supervised vision transformers [6,14] trained on large whole-slide collections have produced general-purpose patch encoders such as UNI [3] and CONCH [10], compared systematically in benchmarks such as PathBench [11]. Of these, H-optimus-1 [2,15] is a large vision transformer trained on a multi-institutional histology corpus, and Virchow2 [21] is a mixed-magnification transformer that augments its class-token representation with pooled spatial tokens. Both yield strong linear-probing performance, motivating their use as fixed encoders here.

Fig.2 gives an overview of the method. Every Haematoxylin and Eosin (H&E) patch is passed through two frozen pathology foundation models and a lightweight classifier head converts each backbone's embedding into class probabilities. The H-optimus-1 probabilities are calibrated per class and the two probability vectors are combined by a fixed weighted average before the final arg-max decision.

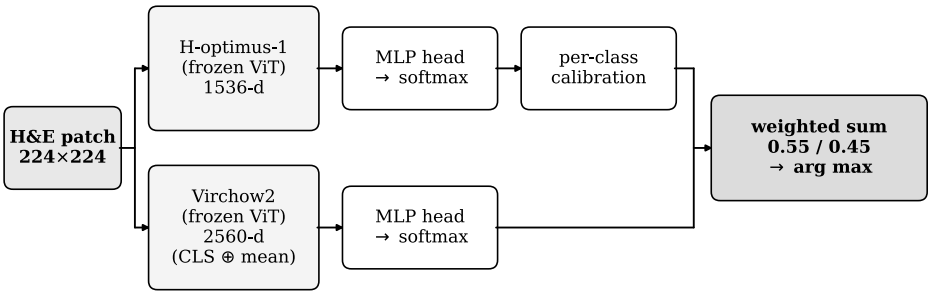

Fig. 2. Overview of the two-backbone ensemble.

The BraTS-Path 2026 dataset comprises more than one million labeled H&E patches drawn from over two hundred digitized tissue sections, together with additional unlabeled whole-slide images for optional self-supervised learning. Labeled data are distributed as WebDataset [17] tar shards. Each labeled sample contains a JPEG patch (90% quality) and a class file with an integer label in the range 0 to 9. A mapping file associates every patch with its patient and slide of origin. The ten classes are highly imbalanced. Frequent categories such as CT and NOTA dominate, while categories such as WM and DM are comparatively rare.

Because patches from the same patient are visually correlated, a naive random split risks optimistic, leaked estimates of generalization. We therefore construct a patient-level split for each class in turn, 25% of the patients contributing that class are drawn (fixed random seed) and added to the held-out set. Because a single patient typically contributes patches of several classes, these per-class draws are accumulated into one patient-level held-out set, and a patient selected on behalf of any class is held out in its entirety. The split is therefore strictly patient-disjoint, and every patch of a held-out

patient is excluded from training regardless of its label. All model-selection and calibration decisions reported below are made on this held-out set.

2.1 Foundation Model Feature Extraction

Image patches are resized to 224×224 pixels and preprocessed using the normalization statistics recommended for each model. Rather than applying stain normalization techniques such as Macenko [12], we retain each model's native preprocessing to preserve consistency with its pre-training distribution. Embeddings are extracted once and cached. H-optimus-1 patches are normalized with the model's published channel statistics (mean 0.707, 0.579, 0.704; standard deviation 0.212, 0.230, 0.178), producing a 1536-dimensional pooled embedding e_h computed in full precision to match the statistics under which its classifier head was trained. Virchow2 patches are normalized with standard ImageNet statistics; following the model's recommended usage, we form a 2560-dimensional descriptor e_v by concatenating the class token with the mean of the K spatial patch tokens (Eq. 1), computed under mixed-precision autocast.

$$e\_v = [\, z\_cls \,;\, (1/K)\, \Sigma\_k\, z\_k \,] \tag{1}$$

2.2 Classifier Heads

For each backbone we train an identical multilayer-perceptron (MLP) head on its embeddings. The head maps the input embedding through two hidden blocks of width 256 and 128 respectively, each consisting of a linear layer, batch normalization, a ReLU nonlinearity, and dropout with rate 0.5, followed by a final linear layer producing ten class logits. A softmax converts the head's logits into a class-probability vector (Eq. 2), where 'e' denotes the embedding of the corresponding backbone and g_θ is the MLP head with learnable parameters θ. The two heads are trained independently, so θ differs between backbones.

$$p = softmax(\, g\_\theta(\, e \,) \,) \tag{2}$$

Strong class imbalance is addressed on two fronts. First, each class c is weighted by inverse frequency (Eq. 3), where N is the number of training patches, C = 10 is the number of classes, and n_c is the number of patches of class c:

$$w\_c = N / (\, C \cdot n\_c \,) \tag{3}$$

Second, the heads are trained with a class-weighted focal loss [8] (focusing parameter γ = 2), which down-weights easy, well-classified patches (Eq. 4):

$$\ell = -\, w\_y\, (\, 1 - p\_y \,)\hat{}\gamma\, \log p\_y \tag{4}$$

where p_y is the predicted probability of the true class y. We further draw minibatches with a weighted random sampler so that under-represented classes appear more often during optimization. The heads are optimized with AdamW (learning rate 5×10^{-4},

weight decay 0.03) under a cosine-annealing schedule, with early stopping on the held-out macro-averaged F1.

2.3 Per-Class Probability Calibration

Arg-max over the raw softmax output of an imbalanced classifier tends to under-predict rare classes. To counter this on the H-optimus-1 head, we scale its probability vector p_h element-wise by a fixed per-class multiplier vector m and renormalize to a distribution q_h before the decision (Eq. 5), where $\odot$ is the element-wise product:

$$q\_h = (m \odot p\_h) / \| m \odot p\_h \|_1 \quad (5)$$

Per-class calibration multipliers were selected via a greedy search that maximizes macro-F1 on the patient-level held-out validation set. Most classes retained their original probability weight. WM received a multiplier of 3.0, DM received 2.0, and PL and NOTA each received 1.5, while CT, IC, LI, MP, NC and PN remained at 1.0.

2.4 Weighted Ensemble

The final prediction combines the two heads by weighted soft-voting, a standard way to exploit complementary models [5,7]. With the calibrated H-optimus-1 distribution q_h and the Virchow2 distribution p_v, the ensemble probability (Eq. 6) and the predicted class (Eq. 7) are:

$$p = 0.55\, q\_h + 0.45\, p\_v \quad (6)$$

$$\hat{y} = \arg\max\_c\, p\_c \quad (7)$$

The 0.55/0.45 weighting was chosen by evaluating $w \in \{0.55, 0.65, 0.80\}$ on the challenge validation leaderboard, of which 0.55 scored highest. Unlike weight-space averaging [18], this decision-level combination keeps both encoders intact. Section 3.4 reports a sensitivity analysis over this weight on the patient-disjoint held-out set.

2.5 Experimental Setup

Embeddings are extracted with the timm library. Inference uses a data loader with shuffling disabled to guarantee deterministic, reproducible outputs. All experiments were run on a single NVIDIA A100 GPU.

3 Results

We evaluate on the official BraTS-Path 2026 validation set using the challenge metrics: F1 score, accuracy, area under the ROC curve (AUC), Matthews correlation coefficient (MCC), sensitivity, and specificity. Model-selection and calibration decisions were made on the patient-level held-out split described in Section 2.

3.1 Validation-Leaderboard Performance

Table 1 reports performance on the official BraTS-Path 2026 validation set for the two individual heads and the weighted ensemble. Fusion improves over both backbones on every metric: macro-F1 rises to 0.570 from 0.555 and 0.538, and the same ordering holds for accuracy, AUC, MCC and sensitivity, while specificity matches the better of the two. The gains are consistent in direction across all six criteria rather than confined to a single favourable metric.

Table 1. Validation-leaderboard performance of the two individual backbones & the ensemble.

Model	F1	Accuracy	AUC	MCC	Sensitivity	Specificity
H-optimus-1 (calibrated)	0.555	0.871	0.917	0.821	0.561	**0.982**
Virchow2	0.538	0.822	0.871	0.758	0.571	0.980
Ensemble (0.55/0.45)	**0.570**	**0.878**	**0.922**	**0.830**	**0.578**	**0.982**

The metric profile is characteristic of a strongly imbalanced multi-class problem. A specificity of 0.982 alongside a sensitivity of 0.578 describes a model that is reliable when it rejects a class but misses a substantial share of positives in the scarcest categories. Because macro-averaging weights every class equally regardless of prevalence, those categories dominate the averaged error even though most patches are classified correctly; the gap between accuracy (0.878) and macro-F1 (0.570) therefore follows from the class distribution rather than from unstable optimisation. The two backbones reach similar macro-F1 but differ in their error patterns, which motivates combining them at the decision level examined directly in Section 3.3.

3.2 Per-Class Performance (Development Held-Out)

The official validation set is unlabelled and our submitted models are trained on all labelled data, so a per-class breakdown cannot be obtained from either without leakage. We therefore retrain both heads on the patient-level training split alone and evaluate the calibrated ensemble on the held-out patients. This development ensemble reaches an accuracy of 0.700 and a macro-F1 of 0.445 for a single representative seed. Table 3 reports the corresponding means over three seeds. Trained on a subset of the data, it is not identical to the leaderboard model, but its per-class profile is informative.

Table 2 shows that performance tracks class frequency closely. The frequent categories NC, CT, IC and NOTA are recognised reliably, while WM, MP, LI and PL remain the bottleneck. Where PL and LI are almost never recovered even with calibration up-weighting them, and since macro-F1 weights all ten classes equally, these two alone account for much of the gap between accuracy and macro-F1. The breakdown also qualifies the benefit of fusion. The ensemble is strongest on most classes, but on several rare ones averaging is actively harmful. The H-optimus-1 alone is better on DM and WM, and Virchow2 alone on MP. In each case a minority-class instance is recognised confidently by one backbone and assigned low probability by the other, so the

linear average dilutes the confident prediction and the arg-max reverts to a majority class. Fusion thus consolidates performance on well-represented categories while eroding it on precisely the rare ones that dominate the macro-average, an asymmetry is examined further in Section 3.4.

Table 2. Per-class F1 on the patient-disjoint held-out set for the two individual heads and the ensemble, for a single representative seed.

Model	CT	DM	IC	LI	MP	NC	PL	PN	WM	NOTA
H-optimus-1	0.781	**0.483**	0.748	0.011	0.179	0.869	**0.002**	0.518	**0.197**	0.720
Virchow2	0.676	0.350	0.746	**0.013**	**0.228**	**0.886**	0.000	0.506	0.107	0.521
Ensemble	**0.786**	0.384	**0.753**	**0.013**	0.205	0.881	0.000	**0.556**	0.144	**0.730**

3.3 Complementarity of the Two Backbones

To test whether the two backbones are complementary rather than redundant, we compare their predictions on the patient-disjoint held-out set (n = 1,176,661). The two heads predict the same class on 76.0% of patches. Partitioning instead by correctness, both are right on 57.9% and both wrong on 24.7%, leaving 17.4% recovered by exactly one model that is 11.2% by H-optimus-1 alone and 6.2% by Virchow2 alone. This last fraction is what any fusion scheme has to work with. A McNemar test on the discordance is highly significant (chi-square = 17,201, $p < 0.001$), so the two heads fail on systematically different patches rather than making interchangeable errors. An oracle that always selected the correct head would reach 0.754 accuracy, against 0.692 and 0.641 individually. Section 3.4 shows how much of that headroom a fixed linear weighting recovers and Fig. 3 shows the corresponding confusion matrices.

3.4 Ablation and Weight Sensitivity

The ablation distinguishes two effects that aggregate reporting conflates. Per-class calibration corrects for class prior imbalance, and its benefit is confined to minority-class retrieval. The sensitivity on the H-optimus-1 head increases by 0.009, exceeding the seed-to-seed standard deviation, while accuracy and MCC are unchanged. Decision-level fusion exploits the complementarity established in Section 3.3, and its benefit falls instead on the aggregate metrics. The accuracy increases from 0.690 to 0.698 and MCC from 0.604 to 0.614 relative to the best single head, in both cases approximately twice the standard deviation. Neither intervention improves macro-F1 beyond seed variation. The two mechanisms are therefore effective but non-overlapping, and the oracle headroom quantified in Section 3.3 is only partially realised by a fixed linear weighting. The submitted configuration is the only one that lies within one standard deviation of the

best on all four metrics. Equal weighting performs comparably to the submitted 0.55/0.45 split, a dependence Fig. 4 examines across the full weight range.

Table 3. Ablation on the patient-disjoint held-out set, reported as mean ± standard deviation over three seeds. Bold marks entries best standard deviation within.

Configuration	macro-F1	Accuracy	MCC	Sensitivity
H-optimus-1, raw	0.429 ± 0.013	0.690 ± 0.004	0.604 ± 0.006	0.460 ± 0.005
H-optimus-1, cal	**0.437 ± 0.012**	0.689 ± 0.005	0.602 ± 0.006	**0.469 ± 0.005**
Virchow2	0.400 ± 0.004	0.645 ± 0.006	0.551 ± 0.007	0.450 ± 0.002
Ens. 0.55/0.45, raw	0.430 ± 0.008	**0.700 ± 0.004**	**0.617 ± 0.005**	0.460 ± 0.002
Ens. 0.55/0.45, cal	0.434 ± 0.010	0.698 ± 0.004	0.614 ± 0.005	0.464 ± 0.004

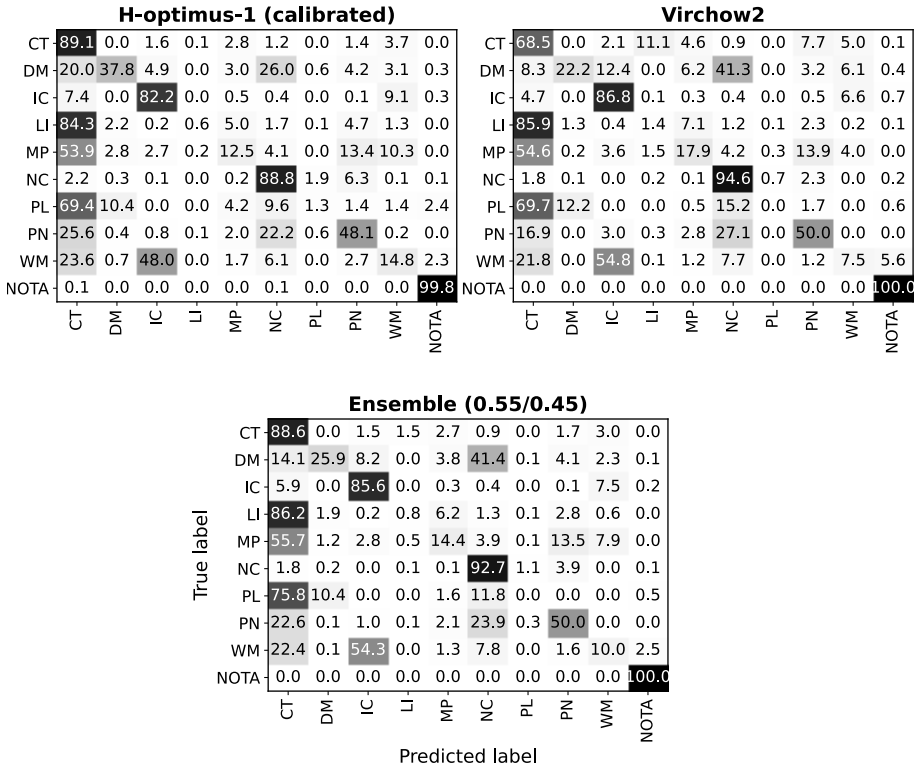

Fig. 3. Row-normalized confusion matrices on the patient-disjoint held-out set for individual models and the combination. Cell values are percentages of each true class, darker shading indicates a higher percentage and the diagonal gives per-class recall, while dark off-diagonal cells mark systematic confusions, most prominently the absorption of LI, PL and MP into CT, and of WM into IC.

Fig. 4 plots macro-F1 against the fusion weight. The curve rises steeply from pure Virchow2 and then flattens, with a maximum of 0.440 at w = 0.75 that exceeds the submitted 0.55 by 0.006. Three considerations argue against adopting that maximum. First, the gap is approximately half a standard deviation, and every weight in [0.50, 1.00] lies within one standard deviation of the peak, so three seeds do not resolve an optimum. Second, seed-to-seed variability is greater in the high-weight region, 0.010 at the submitted weight against 0.013 at the nominal maximum. The apparent gain is accompanied by less stable behavior. Third, and most importantly, this held-out split is the same one on which the calibration multipliers were fitted. Selecting the fusion weight on it as well would compound rather than correct the selection bias acknowledged in Section 4. The weight was chosen from leaderboard feedback, where candidate values on either side of 0.75 were evaluated and scored lower. We therefore retain 0.55 and treat the response as flat rather than peaked.

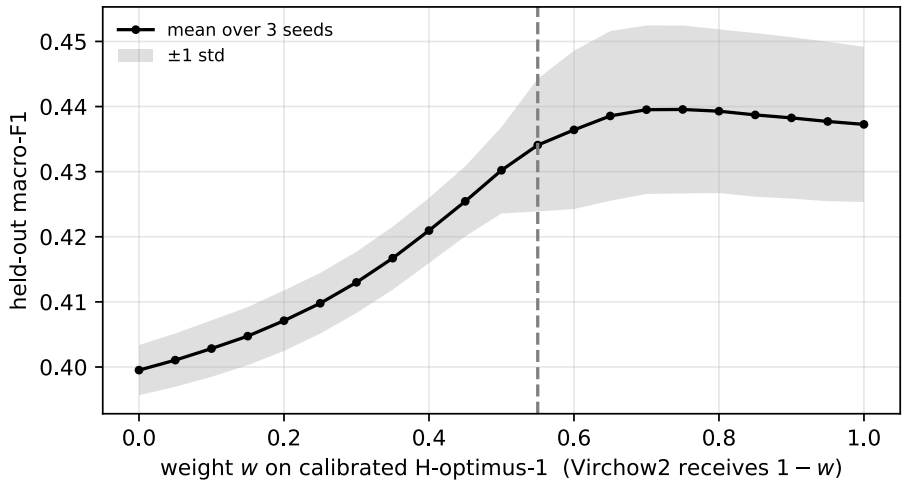

Fig. 4. Held-out macro-F1 as a function of the weight on the calibrated H-optimus-1 head, averaged over three seeds with a ±1 standard deviation band. The dashed line marks the submitted value of 0.55.

4 Discussion

Our work shows that two pathology foundation models, used frozen and combined at the decision level, deliver competitive fine-grained histologic classification at a fraction of the cost of end-to-end training: embeddings are computed once and cached, so every subsequent experiment is cheap enough to repeat across seeds. The weight that appears optimal on a single run proves indistinguishable from the submitted 0.55 once variance is measured. The two encoders are genuinely complementary, with 17.4% of held-out patches recovered by exactly one of them, yet fusion improves accuracy and MCC without moving macro-F1, because a fixed convex combination dilutes the confident

minority-class prediction on which that metric depends and the arg-max reverts to a more abundant class.

Calibration raises minority-class sensitivity while leaving the aggregate metrics untouched, and fusion does the reverse, retaining both is therefore coverage rather than redundancy, and the submitted configuration is the only one of six that remains within a standard deviation of the best on all four metrics. Two caveats apply: the fusion weight was selected from public validation-leaderboard feedback and the calibration multipliers by a greedy search on a single patient-level split, so neither is free of selection bias. Against this, every patch belonging to a held-out patient is excluded from training, so the held-out figures describe performance on patients the models have never encountered rather than unseen crops of familiar slides. The categories that remain hardest to recover are also among the most consequential such as microvascular proliferation is one of the histologic features that establishes grade 4 disease under the current WHO classification [9]; leptomeningeal infiltration bears on the extent of spread; lymphocytic content characterises the immune microenvironment. Failure on these classes is therefore not simply a lower average, but failure on the observations that carry the diagnostic decision. On this evidence the route forward runs not through larger backbones, whose representations are already complementary and demonstrably underexploited, but through fusion that adapts to the class and to the input, evaluated at the level of patients rather than patches.

Acknowledgments. This work used the BraTS-Path 2026 dataset provided via the Synapse platform. We acknowledge the challenge organisers for data access and challenge resources**.**

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bakas, S., et al.: BraTS-Path challenge: assessing heterogeneous histopathologic brain tumor sub-regions. arXiv:2405.10871 (2024)
2. Bioptimus: H-optimus-1: foundation model for histology. https://huggingface.co/bioptimus/H-optimus-1 (2025)
3. Chen, R.J., et al.: Towards a general-purpose foundation model for computational pathology. Nat. Med. 30, 850–862 (2024). doi:10.1038/s41591-024-02857-3
4. Daud, A., Makris, D., Rahman, F.: Efficient classification of glioblastoma sub-regions using MobileNetV2 (2026)
5. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple Classifier Systems. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000). doi:10.1007/3-540-45014-9_1
6. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: ICLR (2021)
7. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS, vol. 30 (2017)
8. Lin, T.-Y., et al.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017). doi:10.1109/ICCV.2017.324

9. Louis, D.N., et al.: The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro-Oncology 23(8), 1231–1251 (2021). doi:10.1093/neuonc/noab106
10. Lu, M.Y., et al.: A visual-language foundation model for computational pathology. Nat. Med. 30, 863–874 (2024). doi:10.1038/s41591-024-02856-4
11. Ma, J., et al.: PathBench: a comprehensive comparison benchmark for pathology foundation models towards precision oncology. arXiv:2505.20202 (2025)
12. Macenko, M., et al.: A method for normalizing histology slides for quantitative analysis. In: IEEE ISBI, pp. 1107–1110 (2009). doi:10.1109/ISBI.2009.5193250
13. Monroy, L.C.R., Mayr, M., Mill, L., Köstler, H., Maier, A.: Patch-level brain tumor subregion classification using foundation models under long-tailed data distributions (2026)
14. Oquab, M., et al.: DINOv2: learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024)
15. Scalbert, M., et al.: H-optimus-1: a foundation model for computational histopathology. Cancer Res. 86(Suppl.), LB174 (2026). doi:10.1158/1538-7445.AM2026-LB174
16. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: ICML, pp. 6105–6114 (2019)
17. WebDataset: a high-performance dataset format for PyTorch. https://pypi.org/project/webdataset (2024)
18. Wortsman, M., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: ICML, pp. 23965–23998 (2022)
19. Zhang, J., Zhong, Q., Weng, Y., Chen, K.: Patch-level glioblastoma subregion classification with a contrastive learning-based encoder (2026)
20. Zhang, Y., et al.: Deep learning for glioblastoma morpho-pathological features identification: a BraTS-Pathology challenge solution. arXiv:2507.18133 (2025)
21. Zimmermann, E., et al.: Virchow2: scaling self-supervised mixed magnification models in pathology. arXiv:2408.00738 (2024)