Pathology Foundation Models Under Class Starvation and Distribution Shift: NeuroForge at the BraTS 2026 Pathology Challenge

Jianfeng Zhang[1,2], Zhiwei Hou[2,3], Mengshan Guan[2], Xinping Huang[2,4], and Hui Zhang[5⋆]

[1] Guangdong Medical University, Zhanjiang, China
[2] Department of Neurosurgery, Sun Yat-sen University Cancer Center, Guangzhou, China
[3] The First Affiliated Hospital of University of Science and Technology of China (Anhui Provincial Hospital), Hefei, China
[4] Zhongshan Longkong Artificial Intelligence Center, Zhongshan, China
[5] Shandong Cancer Hospital and Institute, Jinan, China

Abstract. We describe the NeuroForge submission to the BraTS 2026 pathology challenge (BraTS-Path), which asks for patch-level classification of digitized glioma tissue sections into nine histologic tissue classes plus a "None of the above" (NOTA) class. The task couples two difficulties that interact badly: extreme class imbalance with patient-level concentration of several classes — the NOTA class (15.3% of patches) lives in only 5 of the 126 labeled patients, presence of lymphocytes (PL) in 7, and leptomeningeal infiltration (LI) in 10 — and a distribution shift between the labeled training data and the official validation data, which makes any locally constructed validation split unrepresentative by construction. Our final system is a seven-model ensemble of patch classifiers built on near-frozen pathology foundation model backbones (backbone learning rate 100× lower than the head): two ResNet-34 networks initialized from ImageNet, two phikon ViT-B/16 models (self-supervised iBOT pretraining on 40M histology tiles) and three phikon-v2 ViT-L/16 models (DINOv2 pretraining on 460M tiles), combined by softmax averaging with flip test-time augmentation over a whole-patch input view. Two findings mattered more than any architecture choice. First, a routine patient-level holdout had silently swept 98.8% of one class out of the training pool, driving its recall to exactly zero; retraining on all 126 patients with true-prior class weighting restored the dead class with no collateral damage. Second, architecture diversity decorrelates ensemble errors far more than seed diversity (90–96% versus 98.6% agreement with the ensemble), which justified spending the ensemble budget on heterogeneous backbones. On the public validation leaderboard the ensemble reached an accuracy of 0.8755 (rank 15 of 20 at the July 2026 snapshot). We also report a series of negative results — native-resolution multi-crop

⋆ Corresponding author: Hui Zhang (long.2024@163.com).

evaluation and expectation-maximization prior correction — that follow directly from the distribution shift, and we argue that on this task the leaderboard is the only unbiased validator available to participants.

Keywords: Computational pathology · Glioma histology · Foundation models · Class imbalance · Distribution shift · Deep learning ensembles · BraTS challenge.

1 Introduction

The BraTS 2026 pathology challenge (BraTS-Path) evaluates patch-level tissue classification in digitized glioma sections: given a histology patch, predict one of nine tissue classes (official names in Table 1) — cellular tumor (CT), dense macrophages (DM), infiltration into the cortex (IC), leptomeningeal infiltration (LI), microvascular proliferation (MP), geographic necrosis (NC), presence of lymphocytes (PL), pseudopalisading necrosis (PN), white matter penetration (WM) — or NOTA (none of the above) [1]. Submissions are containerized and scored by the organizers on hidden data [2]; during the validation phase each team may score a limited number of submissions on the public leaderboard. Patch-level glioma tissue typing matters beyond the challenge itself: the relative composition of these tissue classes is a quantitative readout of tumor heterogeneity, and automated pipelines for it must survive exactly the pathologies we study here — rare classes, site shift, and labels that describe whole regions rather than pixels.

Two structural properties of the dataset dominate method design. (i) *Patient-level class concentration*: several classes are present in only a handful of patients — NOTA in 5, PL in 7, LI in 10, DM in 12 of the 126 labeled patients (Table 1). Any patient-level train/validation split therefore risks either starving a class entirely or building a validation fold whose class prior is anti-correlated with the official one. (ii) *Prior shift*: the official validation class prior provably differs from the training prior (NOTA at least 27% versus 15.3% in the released patches), so accuracy numbers computed on any local split are biased estimators of the leaderboard score — in both directions, depending on the split.

Under these conditions we organized development around single-variable probes whose outcome is decidable from one leaderboard submission, and we treated locally computed validation numbers as advisory only. The resulting system is deliberately simple: an ensemble of seven patch classifiers on near-frozen foundation-model backbones, trained with a class-starvation fix, combined by flat softmax averaging. This paper describes the system, the starvation and shift diagnoses, and the negative results that constrained the design.

2 Related Work

Pathology foundation models. Self-supervised pretraining on large histology tile corpora has produced general-purpose feature extractors that transfer

strongly to downstream classification with minimal fine-tuning. Publicly downloadable families include phikon [5] (iBOT [7] pretraining on 40M TCGA tiles) and phikon-v2 [6] (DINOv2 [8] pretraining on 460M tiles); gated families such as UNI [9] and Virchow [10] report stronger downstream performance but require approved access. Our results are consistent with this literature at challenge scale: replacing ImageNet initialization with pathology foundation models gave the single largest improvement we observed, and the gap to the gated models is visible at the top of the leaderboard.

Class imbalance and label scarcity. Standard remedies for long-tailed recognition include inverse-prior loss re-weighting, resampling, and logit adjustment. We use inverse-true-prior weighting because it is the least invasive option that does not duplicate patches across epoch boundaries; its success here depended less on the weighting itself than on the upstream diagnosis that a patient-level split had destroyed the class in the first place — weighting cannot recover a class that the dataloader never sees.

Prior and domain shift adaptation. When the label prior changes between training and deployment, posterior correction by EM or by black-box shift estimation can in principle recalibrate a classifier. Our negative result with EM correction is a cautionary instance of this family: the correction is only as good as the prior it is anchored on, and an anchor estimated from an anti-correlated local split amplifies the bias it is meant to remove.

3 Task, Data and Metric

The released training data comprise 126 patients and 255 digitized tissue sections, from which the organizers released 1,631,432 labeled patches of 512×512 pixels in 40 tar shards (the current challenge release used throughout this work), each labeled with one of the ten classes listed in Table 1 (names and indices per the official task description and `class_map.json`). The official evaluation metrics are the Matthews correlation coefficient (MCC) and the per-class-averaged F1, computed on the hidden validation and testing sets, with accuracy and weighted per-class AUC-ROC additionally published; the container must emit a single `predictions.csv`. Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097). We used only the officially released data and publicly downloadable pre-trained weights; no private or external annotated data were used.

Two dataset properties shaped everything that follows. First, patient-level class concentration: several classes live in a handful of patients (PL in 7, NOTA in 5, LI in 10, DM in 12 of 126), so a random patient-level split can destroy a class's training pool. Second, prior shift: the official validation class prior provably differs from the released one (NOTA at least 27% there versus 15.3% here), so locally computed accuracy is a biased estimator of the leaderboard score.

Table 1. The ten BraTS-Path classes with official names (as listed in the challenge task description and `class_map.json`), patch counts in the released labeled training set (1,631,432 patches, 40 tar shards, 126 patients, 255 slides), and patient coverage. The label describes the dominant content of the whole 512×512 patch.

Abbr.	Official class name	Patches	Patients
CT	Presence of cellular tumor	431,769 (26.5%)	102
DM	Regions with dense macrophages	30,738 (1.9%)	12
IC	Infiltration into the cortex	173,994 (10.7%)	48
LI	Leptomeningeal infiltration	41,326 (2.5%)	10
MP	Areas abundant in microvascular proliferation	83,440 (5.1%)	40
NC	Geographic necrosis	336,179 (20.6%)	69
PL	Presence of lymphocytes	33,031 (2.0%)	7
PN	Pseudopalisading necrosis	143,976 (8.8%)	68
WM	Penetration into white matter	106,979 (6.6%)	39
NOTA	None of the above	250,000 (15.3%)	5

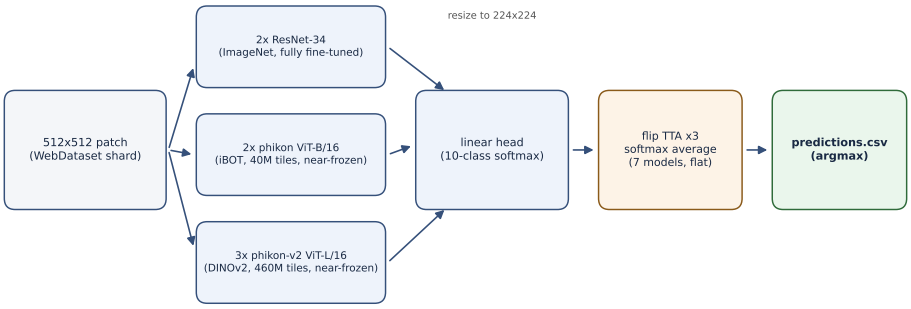

Fig. 1. Pipeline overview. Seven patch classifiers (2× ResNet-34 ImageNet, 2× phikon ViT-B/16, 3× phikon-v2 ViT-L/16) with linear heads; flat softmax averaging with 3-view flip TTA; single `predictions.csv` output.

4 Methods

Figure 1 summarizes the system: a whole 512×512 patch is resized to 224×224 and scored by seven classifiers across three backbone families; their softmax outputs are averaged (flat) with flip TTA, and the argmax forms the prediction.

4.1 Foundation-model ensemble

The final submission averages the softmax outputs of seven patch classifiers: two ResNet-34 [3] models initialized from ImageNet weights, two phikon ViT-B/16 models [5] (iBOT self-supervised pretraining [7] on 40M TCGA histology tiles) and three phikon-v2 ViT-L/16 models [6] (DINOv2 pretraining [8] on 460M histology tiles). Each classifier couples a near-frozen backbone with

a trained linear classification head optimized with cross-entropy and inverse-prior class weights (Sect. 4.2). All models consume the *whole-patch view*: each 512×512 patch is resized to 224×224 — the native input resolution of all three backbone families (ResNet-34, phikon ViT-B/16, phikon-v2 ViT-L/16) — rather than cropped, because the patch label describes the patch as a whole and the BraTS-Path regions of interest can occupy only a small sub-region of the patch. Training uses AdamW (weight decay 5e−2): head learning rate 10^{-3} with a 100× gentler backbone rate (10^{-5} for phikon, 8×10^{-6} for phikon-v2), batch size 96 (phikon) / 64 (phikon-v2), 3–4 epochs; the two ResNet-34 models are fully fine-tuned (batch 192). We note the resulting asymmetry explicitly, since it matters for interpretation: the ResNet-34 members are fully fine-tuned while the foundation-model members are near-frozen, so the pretraining-source comparisons in Table 3 confound backbone and training regime by design — under our submission budget, isolating them would have cost slots we did not have. Inference applies flip test-time augmentation (three views per patch) and a flat, unweighted average over the seven models. Pretraining proved to be the only lever that ever moved the score materially: replacing from-scratch training with ImageNet initialization, and ImageNet with pathology foundation models, accounts for essentially all of our improvement over baseline, while every other knob we tried (native-resolution crops, multi-crop coverage, EM prior correction, ensemble re-weighting) was neutral or net-negative (Sect. 5).

The choice of *near-frozen* backbones deserves a justification, because it is against the current fine-tuning fashion. Three reasons drove it. First, the effective data regime is small: 1.6M patches sounds large, but the patches come from only 126 patients, so patient-level effective sample size is tiny and full fine-tuning of a ViT-L/16 overfits patient identity long before it improves tissue decision boundaries. Second, near-frozen features make the starvation and shift diagnoses of Sect. 4.2 interpretable: with the representation essentially fixed, any change in per-class behavior is attributable to the head and the weighting, not to representation drift. Third, such extractors make the ensemble auditable end-to-end — the seven models share no training dynamics of consequence, so their agreement structure (Table 2) reflects architecture-induced inductive biases rather than correlated optimization noise. We verified empirically that this choice is not the bottleneck: the score gap between our models and the leaderboard leaders is consistent with the published gap between ungated and gated foundation models, not with a frozen-versus-fine-tuned gap.

4.2 Class-starvation diagnosis and fix

Our initial patient-level holdout (seed 1234) produced a model whose recall on the PL class was exactly 0.000. The cause was not model capacity but the split: 98.8% of PL's patches had been swept into the held-out patients (397 of its 33,031 patches remained in training). A per-patient audit (Table 1) shows this is a structural property of the dataset, not of that one seed: PL lives in 7 patients, NOTA in 5, LI in 10 and DM in 12 of the 126, so *any* patient-level split carries this risk for some class. The fix has two parts: (i) train the final models on

all 126 patients, abandoning the patient-level holdout entirely; and (ii) weight the loss by the inverse of the true class prior, so that minority classes are not absorbed by the majority ones. After the fix, PL recovered from 1 to roughly 1,000 correct predictions over the 1.63M released patches, and no class remained dead in the final predictions — the same per-class audit confirms the fix did not silently starve any other concentrated class. This diagnosis also has an irreducible cost: with no representative local holdout possible, final model selection had to default to the public leaderboard, which is why every remaining design decision was organized as a single-variable probe decidable from one submission (Sect. 6).

4.3 Training and inference protocol

The training pipeline is deliberately minimal, which we consider a feature rather than a concession: every component had to justify itself on the leaderboard to survive. (i) *Input*: the whole 512×512 patch, resized to the backbone's native input resolution; no tiling, no color normalization beyond each backbone's published preprocessing. (ii) *Model*: near-frozen foundation backbone ($100\times$ lower backbone learning rate) plus a linear classification head; the two ResNet-34 models are initialized from ImageNet and trained in full. (iii) *Objective*: cross-entropy weighted by the inverse of the true training class prior, on all 126 patients (no holdout; Sect. 4.2). (iv) *Inference*: three flip views per patch, softmax outputs averaged first over views and then over the seven models with uniform weights; the argmax of the averaged posterior is the predicted label. (v) *Selection*: every candidate change was evaluated as a single-variable probe against the protected recipe, one leaderboard submission per change, so no decision ever depended on a difference smaller than the leaderboard's own noise.

4.4 Ensemble diversity analysis

We measured per-model agreement with the ensemble decision on the released patches. Models of different architectures (ResNet vs. ViT-B vs. ViT-L) agree with the ensemble on 90–96% of patches, while same-architecture models trained from different seeds agree at 98.6% (Table 2). Architecture diversity therefore contributes most of the ensemble's error decorrelation, and we spent the ensemble budget on heterogeneous backbones rather than on more seeds of the strongest single architecture. A score-weighted blend (weights proportional to solo official scores) was evaluated alongside the flat average and did not beat it; the flat seven-model ensemble was retained.

4.5 Container compliance

The submitted Docker image runs the seven-model ensemble end-to-end with no network access: it reads the input shards from `/input` and writes a single `/output/predictions.csv`, inside the official compute envelope (a single A6000-class GPU and a 24-hour wall-clock budget, which the ensemble uses a

Table 2. Agreement of individual models with the flat seven-model ensemble decision on the released patches. Cross-architecture disagreement is the ensemble's error-decorrelation source; same-architecture seeds are near-redundant.

Model pair	Agreement with ensemble
ResNet-34 (ImageNet) vs. ensemble	90–96% (cross-architecture range)
phikon ViT-B/16 vs. ensemble	90–96% (cross-architecture range)
phikon-v2 ViT-L/16 vs. ensemble	90–96% (cross-architecture range)
same architecture, different seed	98.6%

Table 3. Official public validation metrics of the main model generations (snapshot July 2026), all read from the organizers' published scoring table. The challenge's official metrics are MCC and per-class-averaged F1; accuracy and weighted AUC-ROC are additionally published.

Generation	Accuracy	MCC	F1 (per-class avg)	AUC-ROC (w.)	Note
v1: ResNet-18, from scratch	0.6876	0.5810	0.2684	0.7844	baseline
v3: ResNet-34, ImageNet init	0.8449	0.7841	0.3977	0.9039	+0.157 acc. from pretraining
v6: phikon ViT-B/16	0.8629	0.8090	0.4793	0.9201	pathology foundation model
flat6: 6-model ensemble	0.8754	0.8269	0.4855	0.9272	
flat7: 7-model ensemble	**0.8755**	**0.8271**	**0.4852**	**0.9275**	

small fraction of). The image reproduces the class distribution of the best validation submission — in particular, no class is dead in the emitted predictions — and the full inference path (backbone weights, preprocessing, TTA and averaging) is pinned so that the container output is deterministic. Source code is available at `https://github.com/B664577/brats2026-flair-viaevum`.

5 Experiments and Results

5.1 Leaderboard progression

Table 3 reports the official public validation numbers of the main model generations. **Every number in this paper is the organizers' official validation figure, read from the published scoring table. The challenge's official metrics are the Matthews correlation coefficient (MCC) and the per-class-averaged F1; accuracy and weighted per-class AUC-ROC are additionally published, and we report all of them.** Three effect sizes stand out on the accuracy axis: pretraining is worth +0.157 accuracy (from-scratch to ImageNet) and a further +0.018 (ImageNet to pathology foundation models) at single-model level; ensembling seven heterogeneous backbones adds a final +0.013 over the strongest single foundation model. The same ordering holds on the official MCC and F1 metrics.

5.2 Case study: anatomy of a dead class

The starvation failure is worth dissecting because it is invisible to aggregate metrics until it is fatal. A patient-level split is the textbook defense against slice leakage in histology, and with 126 patients it looks affordable. But the class in question is concentrated in a handful of patients, so holding out a patient sample removed 98.8% of the class's patches; the surviving 397 patches were then absorbed by the majority classes during training, and the model's recall on the class converged to exactly 0.000 while overall accuracy looked healthy. Two properties made the failure hard to catch: the class is small enough that a zero recall moves aggregate accuracy only marginally, and the class prior in any plausible validation fold is itself unstable, so no local alarm fired. The practical checklist we derived from it: audit class counts *per patient* before choosing a split; audit per-class recall, not only accuracy, after every training; and treat a class that disappears under a split as a property of the dataset, not of the model.

5.3 Effect of the starvation fix

We position the starvation fix as a correctness gate rather than a score gate. Its measurable effects are structural: on the released patches the previously dead class recovered from a single correct prediction to roughly one thousand, and the final container emits no dead class at all. We deliberately did not chase a leaderboard delta for the fix in isolation, because a model with a silently dead class is not a strictly worse version of a correct model — it is an incorrect one, and comparing their accuracies rewards the failure mode. The accuracy progression in Table 3 should therefore be read as conditional on the fix: every entry after the diagnosis was trained on all 126 patients with true-prior weighting.

5.4 Negative results and the validation trap

Table 4 collects the branches that did not transfer. Two deserve emphasis. (i) *Native-resolution evaluation*: classifying native-resolution crops with multi-crop (5-crop) coverage scored 0.7952, well below the resized whole-patch view (0.8449) — the label describes the whole 512-px patch, so cropping both loses context and forces an ad-hoc vote. (ii) *Expectation-maximization (EM) prior correction*: re-estimating the class prior on the unlabeled validation inputs and correcting the posteriors failed, because the official validation prior provably differs from the training prior (NOTA $\geq$ 27% vs. 15.3%) while our local split is anti-correlated with it; the correction was therefore calibrated on a prior that points the wrong way. Local validation misled us three times — on crop strategy, on prior correction and on an early ensemble weighting — before we abandoned it as a decision signal for this task.

6 Discussion

The dominant lesson of this task is that *evaluation design*, not model design, is the bottleneck. When several classes live in a handful of patients and the of-

Table 4. Representative negative or closed branches with their official or locally measured outcomes.

Branch	Evidence	Outcome
Patient-holdout model selection (seed 1234)	98.8% of one class swept out; class recall = 0.000	Fatal split artifact; fixed by full-data training + true-prior weighting
Native-resolution crops + 5-crop voting	0.7952 official	Worse than whole-patch resize (0.8449); closed
EM prior correction	local split anti-correlated with official prior	Failed; closed
Score-weighted ensemble blend	weights $\propto$ solo scores	No gain over flat average; flat kept
Seed expansion of best single model	98.6% intra-architecture agreement	Low diversity value; not pursued

ficial prior differs from the released prior, a local split is not a noisy estimate of the leaderboard — it is a biased one with an unknown sign, and we documented three occasions where it actively pointed the wrong way. Our response was to simplify: freeze a strong, well-understood recipe (near-frozen foundation backbones, whole-patch view, flat ensemble) and spend the submission budget on single-variable probes whose result is interpretable from one score. This is admittedly a conservative strategy; its cost is that genuinely new directions each consume a leaderboard slot to test, and its benefit is that no submitted regression was ever misread as progress. The diversity analysis gives a second, transferable lesson: once backbones are near-frozen foundation models, architecture heterogeneity decorrelates errors much more than seed heterogeneity, so ensemble budget should buy architectures, not seeds.

A third observation concerns where the remaining headroom lies. The progression in Table 3 saturates smoothly: each stronger pretraining source adds less than the previous one, and ensembling adds least of all. This is the signature of a task whose residual errors are no longer model-variance but genuine ambiguity between classes (for example IC versus CT at patch level) plus label-prior shift. Under that diagnosis, further gains are more likely to come from better priors and better shift handling than from larger ensembles — provided, of course, that one has a validator trustworthy enough to tell, which is exactly what this task withholds. We therefore see the combination of "near-frozen public foundation backbones + starvation auditing + leaderboard-decidable probes" as a reproducible template for future class-starved, prior-shifted patch classification challenges, rather than as a task-specific trick.

7 Limitations and Future Work

The top of the leaderboard (0.9235) is occupied by systems built on gated pathology foundation models such as UNI [9] and Virchow [10], whose weights are not publicly downloadable; our ungated phikon/phikon-v2 ceiling appears to be near 0.88 under the current recipe. Three limitations are explicit. First, without a representative local validator, our model selection carries leaderboard overfitting risk that we can bound but not eliminate. Second, the flat ensemble treats all seven models as equally reliable on every class, which is demonstrably false for the rarest classes; per-class ensemble routing is untested. Third, whole-patch resizing discards resolution that may matter for subtle microvascular patterns. A distribution-robust validation protocol for this task — one that is unbiased under patient-level class concentration and prior shift — remains an open problem we consider more valuable than another point of accuracy.

Acknowledgments. Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097). We thank the BraTS 2026 organizing committee and all data-contributing institutions. The phikon and phikon-v2 weights are released by Owkin for research use.

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

References

1. Bakas, S., Thakur, S.P., Faghani, S., Moassefi, M., Baid, U., et al.: BraTS-Path Challenge: assessing heterogeneous histopathologic brain tumor sub-regions. arXiv:2405.10871 (2024)
2. Karargyris, A., Umeton, R., Sheller, M.J., et al.: Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence 5, 799–810 (2023). https://doi.org/10.1038/s42256-023-00652-2
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR 2016, pp. 770–778 (2016)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR 2021 (2021)
5. Filiot, A., et al.: Scaling self-supervised learning for histopathology with masked image modeling (phikon). In: MIDL 2023 / Owkin technical release (2023)
6. Filiot, A., et al.: phikon-v2: a large and public feature extractor for biomarker prediction (DINOv2 ViT-L, 460M histology tiles). Owkin technical release (2024)
7. Zhou, J., et al.: iBOT: Image BERT pre-training with online tokenizer. In: ICLR 2022 (2022)
8. Oquab, M., et al.: DINOv2: learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
9. Chen, R.J., et al.: Towards a general-purpose foundation model for computational pathology (UNI). Nature Medicine 30, 850–862 (2024)
10. Vorontsov, E., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection (Virchow). Nature Medicine 30, 2924–2935 (2024)