Rare-Class Conditional Routing with Foundation Model Representations for Glioma Histologic Sub-region Classification

Shubham Innani[1,2], Suhang You[1,2], and Carla Pitarch-Abaigar[1,2], Dimitrios Makris[3], Spyridon Bakas[1,2,3,4,5,†]

[1] Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
[2] Indiana University Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN, USA
[3] Department of Computer Science, Kingston University, London, UK
[4] Departments of Neurological Surgery; Biostatistics and Health Data Science; Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA
[5] Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA
[†] Corresponding Author: spbakas@iu.edu

Abstract. Accurate patch-level classification of adult diffuse glioma sub-regions in hematoxylin and eosin (H&E)-stained whole-slide images (WSI) is essential for assessing morphologic heterogeneity. The MICCAI BraTS-Path 2026 challenge provides a large-scale, multi-institutional benchmark comprising >3,000,000 patch samples annotated for 10 histopathological sub-regions and 80 non-annotated WSIs, with a severe class imbalance posing a central methodological challenge. To address this problem, we use the >1.6M training samples to systematically develop and evaluate fully-supervised (XGBoost, multilayer perceptron, parameter-efficient Low-Rank Adaptation) and semi-supervised paradigms (Centroid-based pseudo-labeling, Cross Distillation). These paradigms build upon representations from 4 pathology foundation models: H-Optimus-1, UNI2, Virchow2, and a domain-specific DINOv3 ViT-L model that we pre-trained on the TCGA-GBM & TCGA-LGG data collections. We further investigate centroid-based pseudo-labeling of unlabeled patches from the provided WSI for semi-supervised XGBoost retraining. Rare-class sensitivity is addressed via a lightweight class-routing ensemble, yielding the best overall BraTS-Path score (average of Matthews correlation coefficient & macro-F1) of 0.750. Our findings demonstrate that class-conditional routing of foundation models for the rarely represented classes substantially improves F1, while preserving strong global discrimination.

Keywords: Glioma · Heterogeneity · Histopathology · Foundation models · BraTS-Path

1 Introduction

Adult diffuse gliomas (ADG) are the most common primary malignant brain tumors, with a well-recognized vastly heterogeneous morphologic landscape [1–3]. Within a single whole slide image (WSI) of an ADG distinct histopathological sub-regions coexist, representing: cellular tumor (CT), necrosis (NC), pseudopalisading necrosis (PN), microvascular proliferation (MP), infiltration into cortex (IC), white matter (WM), leptomeningeal infiltration (LI), dense macrophages (DM), presence of lymphocytes (PL), and non-tumor area (NOTA). Identifying these sub-regions at the patch level is a prerequisite for automated intra-tumoral heterogeneity quantification and analysis, ultimately improving our understanding of this disease. Recent advances in computational pathology have demonstrated the potential of artificial intelligence (AI) and deep learning methods to extract clinically relevant molecular, diagnostic, and prognostic information directly from routine H&E-stained histopathology slides [4–10].

Recent top-performing methods on BraTS-Path 2025 [11–17] converged on two complementary paradigms: (i) fine-tuning pathology foundation models (FMs) with parameter-efficient Low-Rank Adaptation (LoRA [18]), and (ii) extracting frozen FM representations and training gradient-boosted classifiers over representations [11]. Both paradigms require strategies for class imbalance, with rare-class boosting and per-class threshold calibration reaching state-of-the-art performance. Despite this progress, robust patch-level sub-region classification remains an open problem. The BraTS-Path task has been posed over successive challenge iterations [15], yet the severe under-representation of clinically meaningful rare sub-regions, such as DM, LI, and PL, continues to cap performance. The scoring metric, an equal-weight combination of MCC and macro-F1, exposes a persistent tension: models tuned for global discrimination reach high MCC but collapse rare-class F1, whereas models that recover rare classes give up overall agreement. These gaps motivate a systematic, like-for-like comparison across paradigms and backbones, coupled with a mechanism that assigns each sub-region to the model best suited to predict it.

To address these gaps, our work makes 3 contributions: (1) a systematic cross-paradigm evaluation of 5 training strategies across 3 established pathology FMs (H-Optimus-1 [19], UNI2 [20], Virchow2 [21]) and a domain-specific DINOv3 ViT-L model that we pre-trained on patch samples from The Cancer Genome Atlas (TCGA) glioblastoma (TCGA-GBM) [22] and lower-grade glioma (TCGA-LGG) [23] data collections, enabling a direct comparison between representations from broad pan-cancer pathology FMs and glioma-specific models; (2) a semi-supervised adaptation of Cross Distillation of Multiple Attentions (CDMA) [24] for patch-level classification that leverages unlabeled challenge patches, complemented by a centroid-based pseudo-labeling study examining the reliability and limitations of representation-space confidence under patient-level distribution shift, particularly for rare classes; and (3) a class-conditional hybrid ensemble that combines the strong global discrimination of XGBoost with the improved rare-class sensitivity of a vision transformer (ViT) [25] model based classifier,

Table 1. Training set class distribution. Rare classes comprise <3% of patches.

Region	CT	NC	NOTA	IC	PN	WM	MP	LI	PL	DM
Count	431,800	336,199	250,011	174,003	143,991	106,987	83,448	41,329	33,034	30,739
%	26.5	20.6	15.3	10.7	8.8	6.6	5.1	2.5	2.0	1.9

achieving a better balance between Matthews correlation coefficient (MCC) and macro-F1 than either component alone.

2 Methods

2.1 Data

The MICCAI BraTS-Path 2026 Challenge [15, 26] provides an unprecedented benchmark leveraging >3,000,000 patch samples across training, validation, and testing sets. Specifically, the training set, comprises 1,631,432 labeled samples, each 512×512 pixels at 20× magnification, paired with 80 unlabeled WSIs to facilitate self-/semi-supervised innovations. From these unlabeled WSIs we extracted and used for training an additional 649,670 samples using 50% stride. The challenge's official validation set comprises 114,496 samples, distributed to participants without labels. The training set class distribution is severely imbalanced (Table 1, with rare classes DM (1.9%), LI (2.5%), and PL (2.0%) each comprising fewer than 3% of training patches.

2.2 Foundation Models (FMs)

We evaluate four pathology FMs as frozen or fine-tuned extractors: (i) **H-Optimus-1** [19]: A ViT-g/14 (1536-d) model pre-trained by Bioptimus on over 1 million H&E histology slides via self-supervised learning [27], producing patch representations. (ii) **UNI2-h** [20]: A ViT-H/14 model (1536-d) pre-trained on pathology tiles from over 100K WSIs using DINOv2 with SwiGLU activations and 8 register tokens. (iii) **Virchow2** [21]: A ViT-H/14 model (2560-d representation combining CLS token and register token mean) trained on 3.1M WSIs. (iv) **DINOv3-GBM**: A ViT-L/16 model (1024-d) was pre-trained using DINOv3 [28] exclusively on IDH-wildtype cases from the TCGA-GBM and TCGA-LGG collections. This domain-specific backbone encodes GBM-specific morphological representations not emphasized by pan-pathology FMs, providing a complementary signal for glioma sub-region discrimination.

Our domain-specific DINOv3-GBM backbone was pre-trained on patch samples extracted from the TCGA-GBM and TCGA-LGG data collections, which share an institutional origin with a subset of the slides in the BraTS-Path 2026 challenge. We were not provided with slide-level provenance information for the challenge data, and therefore could not exclude overlapping slides from pre-training. This pre-training corpus was substantially smaller than those of the

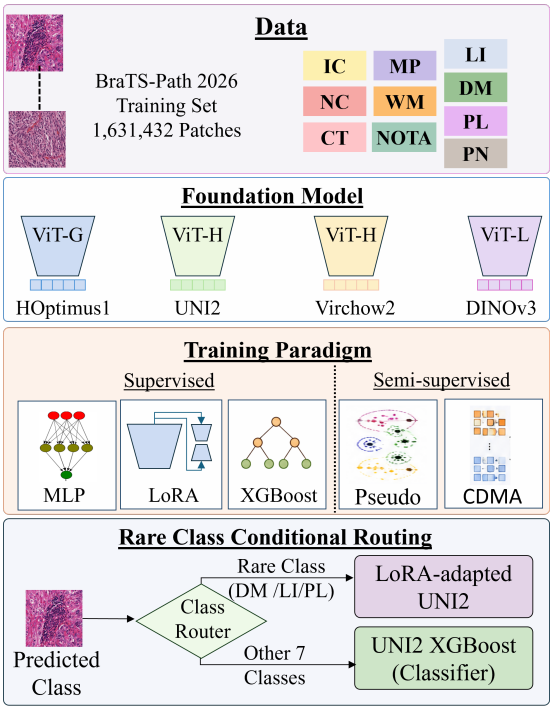

Fig. 1. Overview of the proposed rare-class conditional routing framework for the MICCAI BraTS-Path 2026 challenge.

pan-pathology FMs (∼2,000 WSIs vs. >1M slides for H-Optimus-1, 3.1M for Virchow2), and the DINOv3-GBM model underperformed all three pan-pathology backbones across every training paradigm. The possibility of partial data overlap during pre-training is acknowledged as a limitation. However, given the observed performance gap, potential leakage does not appear to have conferred an empirical advantage.

2.3 Training Paradigms: Supervised Approaches

Figure 1 presents the overall architecture of the proposed framework. **Multilayer Perceptron (MLP)**: Classification head (backbone-dim → 10 classes) is appended to each frozen backbone and trained with cross-entropy loss. Patches are resized to 224×224 and normalized with backbone-specific statistics. We use AdamW with cosine annealing, batch size 256, and 100 training epochs. All configurations used random seed 42 and were trained once.

Low-Rank Adaptation (LoRA [18]): The pre-trained backbone parameters are frozen, and trainable LoRA adapters are inserted into the QKV projections of every transformer block. We use rank $r = 4$, scaling factor $\alpha = 8$, and dropout $p = 0.05$. A single linear classification head maps the backbone representation

to the 10 target classes. Training uses AdamW with a weight decay of 0.05 and separate learning rates for the classification head and the LoRA-adapted backbone parameters. The head is optimized with a learning rate of (5×10^{-5}) and LoRA-adapted backbone parameters (5×10^{-6}). A OneCycleLR schedule with a 5% warm-up is applied for up to 50 epochs. Early stopping with a patience of 15 is applied after a minimum of 50 epochs (i.e., no stopping before epoch 50). Batch sizes are 48 for UNI2 and 32 for Virchow2 and H-Optimus-1. The objective is weighted cross-entropy with inverse-frequency class weights normalized to sum to the number of classes and label smoothing of 0.1. Training augmentations include random horizontal and vertical flips, rotations in multiples of 90°, and color jitter with brightness, contrast, and saturation factors of 0.15 and hue factor of 0.05. Random seed 42 was used for all runs and each configuration was trained once and evaluated on the official Synapse leaderboard.

XGBoost: Frozen representations are independently extracted using H-Optimus-1, UNI2, and Virchow2, producing representations of 1,536, 1,536, and 2,560 dimensions, respectively. For each backbone, the training samples are divided into 10 stratified subsets. Samples from relatively common classes are distributed across the 10 subsets, whereas samples from the designated rare classes (DM, LI, MP, PL, and WM) are included in every subset. Each subset is used to train an independent XGBoost classifier [29], and the resulting 10 class-probability vectors are averaged at inference. Each classifier uses 300 trees, a maximum depth of 6, a learning rate of 0.05, a subsample ratio of 0.8, a column-sampling ratio of 0.4, a minimum child weight of 5, and $\gamma = 0.1$. For each subset, sample weights are set to the inverse frequency of each class within that subset and normalized so that the mean sample weight equals 1. After averaging probabilities, a per-class rare-boost multiplier $(\max\_count/\mathrm{count}_c)^{0.3}$ is applied and the result is renormalized. Per-class decision thresholds θ_c are then optimized by coordinate ascent (3 rounds, 30 grid points per class in $[0.2, 5.0]$) to maximize macro-MCC on the internal training-derived validation split; the final prediction is $\arg\max_c P(c)/\theta_c$. Training uses the multi-objective, multinomial log-loss for evaluation, and the histogram tree-building algorithm.

2.4 Training Paradigms: Semisupervised Approaches

Cross Distillation based approach: We adapt CDMA [24] to the patch classification setting. Three independent classification heads operate on the same backbone representation. Each head consists of a linear layer, a GELU activation, dropout with rate 0.5, and a final linear classifier. To encourage diversity among the three branches, the first head receives the backbone representation directly, whereas multiplicative noise sampled from $\mathcal{U}[-0.3, 0.3]$ is applied element-wise to the representations provided to the second and third heads during training only. At inference, no noise is applied, and the three softmax outputs are averaged. Cross knowledge distillation trains each head to match the soft predictions of the other two heads via symmetric pairwise Kullback-Leibler (KL) divergence [30] across all six ordered head pairs, computed on both labeled and unlabeled samples, mitigating the overconfidence that arises from hard pseudo-labels. CDMA

uses the same underlying FM backbone and labeled training set as the MLP baseline, but differs in its multi-head classifier, contrastive projection head, training losses, and use of unlabeled data.

The contrastive component combines supervised contrastive learning on labeled samples with NT-Xent [31] learning on two independently augmented views of each unlabeled image. A shared projection head maps the backbone representation through a two-layer MLP from the backbone dimension to 256 and then to 128. For labeled data, samples sharing a ground-truth class form positive pairs; for unlabeled data, the two augmented views of the same patch form a positive pair, and all remaining samples in the batch act as negatives. Both contrastive losses use a temperature of 0.07. Each labeled minibatch is paired with one unlabeled minibatch of size 64. Optimization uses AdamW with weight decay 0.05, learning rates of 5×10^{-5} for the heads and 5×10^{-6} for the adapted backbone parameters, and a OneCycle learning-rate schedule. At inference, the three softmax probability vectors are averaged, and the final class is selected by arg max.

Centroid-based pseudo-labeling: To exploit the large pool of unlabeled data (∼650K across the provided unlabeled WSIs), we investigated a centroid-based pseudo-labeling strategy for XGBoost retraining. We focused this analysis on XGBoost because it achieved the highest MCC among the evaluated single-model approaches. UNI2 representations were extracted for all unlabeled patches, and class-specific centroids were computed as the L2-normalized arithmetic mean of the training representations assigned to each class. An unlabeled patch received a pseudo-label when its cosine distance to the nearest class centroid fell within the N^{th} percentile of the corresponding class-specific distribution of training-sample distances. This procedure yielded an adaptive threshold for each class, accounting for differences in representation-space density between common and rare classes. We evaluated thresholds at the 35th, 60th, and 80th percentiles, retaining approximately 151K, 334K, and 512K pseudo-labeled patches, respectively.

2.5 Ensemble Strategies

We evaluated several prediction-level ensemble strategies to combine the complementary strengths of neural networks (MLP, LoRA) and XGBoost classifiers. These included probability averaging across backbone-specific XGBoost models, majority voting across LoRA-adapted neural models, and class-conditional hybrid combinations in which predictions for selected rare classes were taken from a neural classifier while the remaining predictions were retained from an XGBoost model. We also evaluated single-model routing combinations using a neural classifier as a rare-class specialist and an XGBoost classifier as the default model for the remaining classes. The rare-classes were selected from LoRA-UNI2, CDMA-UNI2, CDMA-H-Optimus-1, LoRA-H-Optimus-1, and the common-class classifier was selected from XGBoost-UNI2, XGBoost-H-Optimus-1, XGBoost-all 3 model. For these experiments, predictions of DM, LI, or PL from the rare-

class model overrode the corresponding common-class model prediction; all other patches followed the XGBoost classifier.

The routing class set {DM, LI, PL} was initially identified on an internal validation split (10% of training data, stratified by patient ID). We used a 90/10 patch-level internal validation split with patient-aware stratification; because patches from the same patient could occur in both partitions, this split was not patient-disjoint. DM, LI, and PL were the three classes for which the neural classifier consistently recovered sensitivity relative to XGBoost on this internal split. However, because the training data contains multiple patches per patient slide, patient-level disjointness in the internal split could not be guaranteed. Accordingly, these results were treated as directional evidence rather than a reliable held-out estimate. All 12 pairing combinations were subsequently submitted to the official Synapse leaderboard as the primary evaluation. The best configuration (02) was identified post-hoc from the 12 submitted results. We acknowledge that this constitutes implicit selection on the validation leaderboard. However, the same three rare classes were consistently identified as the routing targets in the internal evaluation, suggesting the finding is not purely an artifact of leaderboard selection.

2.6 Evaluation

The BraTS-Path challenge evaluates submitted predictions using six classification metrics: accuracy, area under the receiver operating characteristic curve (AUROC), F1 score, MCC, sensitivity, and specificity. Accuracy measures the proportion of all patches that are correctly classified. AUROC assesses the ability of the model to discriminate between classes across different decision thresholds. The F1 score summarizes the balance between precision and sensitivity, while MCC measures agreement between predicted and ground-truth labels using all entries of the multiclass confusion matrix and is particularly informative for imbalanced classification problems. Sensitivity measures the proportion of ground-truth samples that are correctly identified, whereas Specificity measures the proportion of samples outside a given class that are correctly rejected. Although all six metrics are provided via the Synapse leaderboard, participating methods are officially ranked using the BraTS-Path score (BPS):

$$\mathrm{BPS} = (0.5 \times \mathrm{MCC}) + (0.5 \times \mathrm{F1}) \tag{1}$$

The BPS assigns equal weight to global multiclass discrimination, represented by MCC, and class-balanced predictive performance, represented by the F1 score. All validation-set predictions were evaluated using the official challenge evaluation pipeline on the Synapse leaderboard. The Synapse leaderboard returns only aggregate metrics (accuracy, MCC, AUROC, macro-F1, macro-sensitivity, macro-specificity), and per-class precision, recall, or confusion matrices are not accessible through the challenge evaluation interface. Consequently, class-level performance for individual sub-regions cannot be reported from leaderboard submissions.

Table 2. Challenge validation leaderboard results. CDMA uses the additional 649,670 unlabeled challenge patches and hence is not directly data-matched to the supervised methods.

Backbone	Method	Acc	MCC	AUROC	F1	Sens	Spec	BPS
UNI2	MLP	0.834	0.778	0.890	0.526	0.595	0.988	0.652
	LoRA	0.871	0.822	0.920	0.580	0.615	0.987	0.701
	XGBoost	0.889	0.847	0.938	0.475	0.466	0.977	0.661
	CDMA	0.825	0.762	0.891	0.460	0.482	0.978	0.611
Virchow2	MLP	0.809	0.744	0.867	0.527	0.608	0.986	0.636
	LoRA	0.842	0.781	0.896	0.477	0.496	0.980	0.629
	XGBoost	0.896	0.856	0.939	0.497	0.486	0.978	0.677
	CDMA	0.826	0.764	0.882	0.545	0.598	0.983	0.655
H-Optimus-1	MLP	0.851	0.797	0.898	0.568	0.624	0.989	0.683
	LoRA	0.877	0.828	0.923	0.478	0.483	0.981	0.653
	XGBoost	0.893	0.852	0.938	0.466	0.448	0.977	0.659
	CDMA	0.853	0.800	0.899	0.573	0.630	0.988	0.687
DINOv3-GBM	MLP	0.753	0.674	0.832	0.404	0.483	0.984	0.539
	LoRA	0.723	0.647	0.810	0.425	0.523	0.988	0.536
	XGBoost	0.861	0.807	0.913	0.454	0.453	0.975	0.631
	CDMA	0.761	0.683	0.840	0.405	0.468	0.981	0.544
Ensemble	XGB (3 models)	**0.903**	**0.865**	**0.944**	0.500	0.489	0.980	0.683
	LoRA (3 models)	0.883	0.838	0.925	0.607	0.630	0.986	0.723
	Hybrid (vote)	0.896	0.856	0.932	0.626	0.649	0.987	0.741
	Routing (02)*	0.897	0.858	0.939	**0.641**	**0.651**	0.984	**0.750**

3 Results

Table 2 reports all results on the Synapse BraTS-Path 2026 challenge leaderboard on 114,496 validation patches from independent held-out data.

Key observations. Across the individual backbones, XGBoost consistently achieved the highest MCC, reaching 0.847 for UNI2, 0.856 for Virchow2, and 0.852 for H-Optimus-1, outperforming all neural methods (MLP, LoRA, CDMA) in global discrimination. However, for each of these pan-pathology FMs, XGBoost also produced the lowest macro-sensitivity among all methods and, consequently, a lower macro-F1 than the best-performing neural method for that backbone (UNI2: 0.475 vs. 0.580 with LoRA; Virchow2: 0.497 vs. 0.545 with CDMA; H-Optimus-1: 0.466 vs. 0.573 with CDMA). This confirms that, despite their strong global discrimination, the frozen-representation XGBoost models were the least sensitive to rare classes. LoRA improved MCC over the supervised MLP baseline for all three large pan-pathology FMs, with gains ranging from 0.031 (H-Optimus-1) to 0.044 (UNI2). The largest improvement was observed for UNI2, for which LoRA increased MCC by 0.044 and macro-F1 by 0.054. This suggests that task-specific adaptation of the attention projections is particularly beneficial for UNI2. Since CDMA additionally uses 649,670 unlabeled challenge patches, comparisons with MLP, LoRA, and XGBoost should not be interpreted as controlled architecture-only comparisons.

The domain-specific DINOv3-GBM model underperformed the 3 pan-pathology FMs across all training paradigms. Its strongest result was obtained with XG-

Boost, which achieved an MCC of 0.807, approximately 0.04-0.05 below the corresponding XGBoost results for UNI2, Virchow2, and H-Optimus-1. In addition, LoRA reduced MCC by 0.027 relative to the MLP baseline for DINOv3-GBM, whereas LoRA consistently improved MCC for the larger pan-pathology models. This finding indicates that domain-specific pre-training alone did not compensate for differences in model scale, representation quality, or pre-training diversity. We hypothesize that the underperformance is primarily attributable to pre-training scale: the DINOv3-GBM corpus comprised approximately 2,000 WSIs from TCGA-GBM and TCGA-LGG, substantially smaller than the datasets used to train the pan-pathology FMs (>1M slides for H-Optimus-1, 3.1M WSIs for Virchow2). At this pre-training scale, the breadth of data diversity available to pan-pathology FMs plausibly outweighs the benefit of glioma-specific morphological encoding, consistent with scaling observations in computational pathology [20]. However, this remains a hypothesis rather than a demonstrated finding. Investigation of domain-specific pre-training at larger scale is ongoing work.

Among the ensemble approaches, averaging the predictions of the three independently trained XGBoost models yielded the highest MCC (0.865) and AUROC (0.944). However, its macro-F1 remained limited to 0.500, reflecting continued suppression of rare-class predictions. The three neural network models voting ensemble showed the opposite behavior: macro-F1 increased to 0.607, representing a gain of 0.107 over the XGBoost ensemble, while MCC decreased by 0.027. This trade-off confirms that the ViT models were more responsive to rare classes but less robust in overall discrimination.

The vote-based hybrid provided a more balanced combination of these complementary behaviors. Relative to the three-model XGBoost ensemble, it increased macro-F1 by 0.126 while reducing MCC by only 0.009, resulting in a BPS of 0.741. Although this configuration outperformed both of its individual components, it was subsequently surpassed by the best single-model class-routing configuration described in Section 3.1.

3.1 Class-Routing Sweep

Table 3 reports the single-model routing pairings. The best configuration (02) routes rare classes {DM, LI, PL} to a LoRA-adapted UNI2 model and all remaining classes to a UNI2 XGBoost classifier, reaching a BPS of 0.750 (MCC=0.858, F1=0.641). This single-model pairing outperforms the vote-based hybrid (0.741) while dispensing with both the ViT model majority voting ensemble and the three-model XGBoost, indicating that the routing gain derives from class conditional specialization rather than sub-model ensembling. Pairing the same LoRA-UNI2 rare specialist with an H-Optimus-1 XGBoost (03) is competitive (0.747) and yields the highest routing MCC (0.862). CDMA-based rare specialists (06, 09) and the H-Optimus-1 LoRA rare specialist (11) trail, tracking the lower standalone rare-class F1 of those backbones. Substituting the rare-boosted LoRA-UNI2 variant (04, $^{+}$ = 4 rare classes: DM, LI, MP, PL) degrades both MCC and F1 relative to the plain LoRA-UNI2 specialist. Expanding to a 5-class override

Table 3. Class-routing sweep. Rare classes {DM, LI, PL} predicted by neural classifier; all other classes by common-class model. Rows 04 and 12 use LoRA-UNI2^{+} (4 rare classes: DM, LI, MP, PL); row 13 UNI2^{++} uses a 5-class override set {DM, LI, MP, PL, WM}.

#	Rare-class model	Common-class model	Acc	MCC	AUROC	F1	Sens.	Spec.	BPS
01	LoRA-UNI2	XGB-3model	0.832	0.779	0.898	0.604	0.695	0.993	0.692
02	**LoRA-UNI2**	**XGB-UNI2**	0.897	0.858	0.939	**0.641**	0.651	0.984	**0.750**
03	LoRA-UNI2	XGB-HOpt1	0.901	**0.862**	0.939	0.632	0.634	0.983	0.747
04	LoRA-UNI2^{+}	XGB-UNI2	0.868	0.818	0.921	0.584	0.615	0.985	0.701
05	CDMA-UNI2	XGB-3model	0.817	0.758	0.894	0.483	0.549	0.990	0.621
06	CDMA-UNI2	XGB-UNI2	0.884	0.840	0.935	0.521	0.505	0.979	0.680
07	CDMA-UNI2	XGB-HOpt1	0.887	0.844	0.935	0.513	0.488	0.978	0.678
08	CDMA-HOpt1	XGB-3model	0.817	0.759	0.889	0.548	0.604	0.991	0.654
09	CDMA-HOpt1	XGB-HOpt1	0.887	0.842	0.929	0.579	0.546	0.981	0.711
10	LoRA-HOpt1	XGB-HOpt1	0.894	0.853	0.938	0.530	0.490	0.978	0.692
11	LoRA-HOpt1	XGB-3model	0.824	0.766	0.896	0.501	0.549	0.990	0.634
12	LoRA-UNI2^{+}	XGB-HOpt1	0.874	0.825	0.922	0.584	0.614	0.986	0.705
13	LoRA-UNI2^{++}	XGB-UNI2	0.871	0.823	0.914	0.598	0.621	0.984	0.711

(UNI2^{++}: DM, LI, MP, PL, WM) similarly degrades MCC (0.823 vs. 0.858) and F1 (0.598 vs. 0.641), confirming that routing MP and WM, classes for which XGBoost already performs adequately, to the neural specialist introduces more errors than it corrects.

3.2 Semi-Supervised Pseudo-Labeling

Table 4 reports the pseudo-labeled XGBoost configurations on the challenge leaderboard. All three confidence levels degraded performance relative to the UNI2-only XGBoost baseline (BPS=0.661) and the three-model XGBoost ensemble (BPS=0.683). The p35 model reached MCC=0.820 and F1=0.465 (BPS=0.643), the p60 model MCC=0.802 and F1=0.468 (BPS=0.635), and the p80 model MCC=0.783 and F1=0.462, with BPS monotonically declining. This pseudo-labeling study was exploratory and we include it as a negative result to document that centroid-proximity confidence scoring does not recover performance in this setting. The consistent decline, at minimum, provides evidence that relaxing the threshold does not help, regardless of the precise mechanism. The degradation is consistent with, though not definitively demonstrated to be caused by, class imbalance amplification: rare classes such as LI receive near-zero pseudo-labels even at the lowest threshold (p35), while more morphologically distinct classes attract a disproportionate share of pseudo-labels, diluting the rare-class signal. Class-level pseudo-label accuracy would require access to unlabeled ground truth, which is not available within the challenge framework; accordingly, the mechanistic interpretation is acknowledged as suggestive rather than demonstrated.

Table 4. Semi-supervised pseudo-labeling with UNI2 representations. p_N denotes the percentile confidence threshold for pseudo-label acceptance, accepting approximately 151K (p35), 334K (p60), and 512K (p80) pseudo-labeled patches.

Config	Total pseudo-labels	Acc	MCC	AUROC	F1	Sens	Spec
p35	151,275	0.870	0.820	0.918	0.465	0.503	0.982
p60	333,715	0.854	0.802	0.903	0.468	0.512	0.985
p80	512,186	0.838	0.783	0.890	0.462	0.510	0.986

4 Discussion

We presented a systematic comparison of training paradigms and FMs for glioma sub-region classification from H&E patches, evaluated on the MICCAI BraTS-Path 2026 challenge leaderboard. Our key findings are: (1) Probability-averaged XGBoost models trained on frozen representations from multiple foundation models achieve the highest MCC under distribution shift; (2) LoRA fine-tuning improves over linear probing on MCC across all large backbones; (3) semi-supervised CDMA improves F1 by leveraging unlabeled patches, whereas hard centroid pseudo-labeling degrades under the same distribution shift; and (4) a class-routing ensemble that directs rare classes {DM, LI, PL} to a LoRA-adapted UNI2 model and all other classes to a UNI2 XGBoost classifier achieves the best BPS of 0.750 (MCC=0.858, F1=0.641), improving on a hand-designed vote-ensemble hybrid (0.741) with a markedly simpler single-model design.

Our results reveal a consistent trade-off between MCC and macro-F1 across the evaluated training paradigms. Traditional machine learning (ML) based XGBoost classifiers trained on frozen FM representations achieved strong global discrimination, whereas the newer approaches MLP, LoRA, and CDMA showed greater sensitivity to rare classes. This difference suggests that task-specific optimization of newer methods can better preserve rare-class decision boundaries. The higher MCC performance of XGBoost may be attributed to the use of fixed representations learned from large and diverse histopathology datasets. By retaining the original FM representations, the XGBoost pipelines may preserve more general tissue representations and therefore remain less sensitive to the distribution shift between the pretraining and challenge cohorts.

The effect of semi-supervised CDMA depended strongly on the underlying backbone. For H-Optimus-1, CDMA improved macro-F1 relative to LoRA while maintaining comparable MCC to the supervised MLP baseline. One possible explanation is that cross-distillation among multiple classification heads acts as a form of soft consistency regularization. Rather than relying on a single hard pseudo-label, each head is encouraged to match the softened predictions of the other heads on unlabeled patches. This may reduce collapse toward majority-class predictions and preserve uncertainty for morphologically ambiguous rare classes. Nevertheless, the weaker CDMA results obtained with UNI2 indicate that the benefit of this regularization is backbone-dependent. The contrast between CDMA and centroid-based pseudo-labeling further highlights the impor-

tance of how unlabeled data are incorporated. Both approaches use unlabeled challenge patches, but centroid pseudo-labeling assigns a hard class label according to representation-space proximity. Such assignments are sensitive to class overlap, unequal cluster density, and distribution shift between labeled and unlabeled cohorts.

The ensemble experiments demonstrate that the newer approach (LoRA) combined with traditional XGBoost classifiers captures complementary aspects of the task. The XGBoost ensemble retained strong majority-class discrimination, whereas LoRA achieved a higher F1 score. Instead of averaging probability outputs from models which produced fewer rare-class predictions, the hybrid approach used a class-conditional override rule. Under this rule, the XGBoost prediction was retained by default, while the neural prediction was adopted only for selected rare classes. The routing sweep sharpens this view: UNI2 backbone with LoRA and XGBoost exceeds the routing benefit without any sub-model ensembling, suggesting that the gain is attributable to class-conditional specialization rather than ensemble variance reduction.

Several limitations should be considered. First, the rare-class override sets were selected through empirical inspection of prediction behavior rather than learned using a dedicated gating model. A confidence-aware or validation-optimized routing mechanism may provide a more principled alternative. Second, the domain-specific DINOv3-GBM backbone did not outperform the larger pan-pathology FMs, despite being pre-trained exclusively on TCGA-GBM and TCGA-LGG data that partially overlap with the source cohorts of the BraTS-Path challenge, suggesting that domain specificity alone is insufficient to guarantee superior downstream representations. The extent of slide-level overlap is unknown, and this potential leakage is acknowledged, though the observed performance gap suggests it did not confer an empirical advantage. Third, the Synapse leaderboard does not provide per-class metrics, repeated-run variance estimates, or confidence intervals; each configuration was evaluated as a single submission, and training-run variance across random seeds remains uncharacterized. Finally, the use of unlabeled challenge patches introduces a transductive component that may limit direct comparison with strictly inductive methods.

In conclusion, our rare-class conditional routing framework provides a practical approach for balancing global discrimination with sensitivity to under-represented histopathological sub-regions in ADG. More robust identification of these heterogeneous tissue compartments at the patch level could support downstream computational modeling of spatial and morphological tumor heterogeneity, enabling more detailed characterization of the ADG microenvironment. Such approaches may ultimately facilitate improved understanding of disease biology and, with further validation, contribute to the development of clinically relevant biomarkers and improved patient stratification and outcomes.

Acknowledgments. Computational resources used in this research were partially supported by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. Code is available at `https://github.com/IUCompPath/bratspath2026`

References

1. Sottoriva, A., et al.: Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proceedings of the National Academy of Sciences **110**(10), 4009–4014 (2013)
2. Brennan, C.W., et al.: The somatic genomic landscape of glioblastoma. Cell **155**(2), 462–477 (2013)
3. 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)
4. Innani, S., et al.: PATH-67. AI-Based WHO 2021 Classification of Adult-Type Diffuse Glioma Solely from H&E-Stained Slides. Neuro-Oncology **27**(Supplement_5), v256–v256 (2025)
5. Collins, K., et al.: Artificial Intelligence–Based Classification of Renal Oncocytic Neoplasms: Advancing From a 2-Class Model of Renal Oncocytoma and Low-Grade Oncocytic Tumor to a 3-Class Model Including Chromophobe Renal Cell Carcinoma. Archives of Pathology & Laboratory Medicine **149**(10), 922–929 (2025)
6. Innani, S., et al.: PATH-69. AI-Based Prognostic Risk Stratification of Adult-Type Diffuse Gliomas Using H&E-Stained Slides. Neuro-Oncology **27**(Supplement_5), v257–v257 (2025)
7. Innani, S., et al.: Artificial Intelligence Predicts 2021 WHO Glioma Subtypes from Whole Slide Images. Cancer Research **85**(8_Supplement_1), 6247–6247 (2025)
8. Innani, S., et al.: Interpretable Artificial Intelligence Based Determination of Glioma IDH Mutation Status Directly from Histology Slides. Neuro-Oncology Advances **7**(1), vdaf140 (2025)
9. Innani, S., et al.: AI-Driven WHO 2021 Classification of Gliomas Based Only on H&E-Stained Slides. Neuro-Oncology **28**(1), 282–296 (2026)
10. Innani, S., et al.: PATH-39. AI-Based Identification of Glioma IDH Mutational Status from H&E-Stained Whole Slide Images. Neuro-Oncology **26**(Suppl 8), viii187 (2024)
11. Monroy, L.C.R., et al.: Patch-Level Brain Tumor Sub-region Classification Using Foundation Models Under Long-Tailed Data Distributions. In: MICCAI 2025, LNCS, vol. 16377, pp. 213–222. Springer (2026)
12. Son, J., Roh, T.H.: Patch-Level Classification of Histopathological Subregions in Glioblastoma. In: MICCAI 2025, LNCS, vol. 16377, pp. 183–192. Springer (2026)
13. Zhang, J., et al.: Patch-Level Glioblastoma Subregion Classification with a Contrastive Learning-Based Encoder. In: MICCAI 2025, LNCS, vol. 16377, pp. 193–202. Springer (2026)
14. Arora, M., et al.: GAMMA-Net: Gated Attention Multi-scale Modular Architecture. In: MICCAI 2025, LNCS, vol. 16377, pp. 223–233. Springer (2026)
15. Bakas, S., et al.: BraTS-Path Challenge: Assessing Heterogeneous Histopathologic Brain Tumor Sub-Regions. arXiv preprint arXiv:2405.10871 (2024)
16. Thakur, S., et al.: IMG-121. BraTS-Pathology 2024: Insights and Future Directions Informed by the AI-RANO & RANO-RGP Effort to Assess Glioblastoma Heterogeneity. Neuro-Oncology **27**(Supplement_5), v303–v304 (2025)
17. Thakur, S., et al.: TMIC-60. BRATS-PATH: Assessing Heterogeneous Histopathologic Regions in Glioblastoma. Neuro-Oncology **26**(Supplement_8), viii312–viii312 (2024)
18. Hu, E.J., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR 2022 (2022)

19. Scalbert, M., et al.: Abstract LB174: H-optimus-1: A foundation model for computational histopathology. Cancer Research **86**(8_Supplement), LB174–LB174 (2026)
20. Chen, R.J., et al.: A General-Purpose Self-Supervised Model for Computational Pathology. Nature Medicine **30**, 2057–2070 (2024)
21. Zimmermann, E., et al.: Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology. arXiv:2408.00738 (2024)
22. Scarpace, L., et al.: Radiology Data from The Cancer Genome Atlas Glioblastoma Multiforme (TCGA-GBM) Collection. The Cancer Imaging Archive (2016)
23. Pedano, N., et al.: Radiology Data from The Cancer Genome Atlas Low Grade Glioma (TCGA-LGG) Collection. The Cancer Imaging Archive (2016)
24. Zhong, L., et al.: Semi-supervised Pathological Image Segmentation via Cross Distillation of Multiple Attentions. In: MICCAI 2023, LNCS, pp. 570–579. Springer (2023)
25. Dosovitskiy, A., et al.: An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In: ICLR 2021 (2021)
26. Bakas, S., et al.: BraTS-Path 2026 Challenge: Patch-Level Histopathological Subregion Classification of Diffuse Gliomas. `https://www.synapse.org/` (2026)
27. Filiot, A., et al.: Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. MedRxiv (2023)
28. Simeoni, O., et al.: DINOv3. arXiv:2508.10104 (2025)
29. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: KDD 2016, pp. 785–794 (2016)
30. Joyce, J.M.: Kullback-Leibler divergence. In: International Encyclopedia of Statistical Science, pp. 1307–1309. Springer, Berlin, Heidelberg (2025)
31. Ågren, W.: The NT-Xent loss upper bound. arXiv preprint arXiv:2205.03169 (2022)