This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Neuropathology in the embedding space of glioblastoma heterogeneity: identifying classification-relevant signals with machine learning

Agata Polejowska[1] and Aleksandra Czapiewska[2]

[1] Department of Pathology, Radboudumc, Nijmegen, Netherlands
[2] Department of Clinical Pathomorphology, University Clinical Centre, Gdańsk, Poland

Abstract. Glioblastoma contains several histological tissue patterns that may coexist within the same tumour. Automated recognition of these patterns in digitised histology could support scalable analysis of tumour heterogeneity. In this work, a machine-learning approach was developed and evaluated for automated multi-class classification of regions of interest. Nine computational pathology foundation models were evaluated using patient-level cross-validation. Virchow2 achieved the highest accuracy, whereas GenBio-PathFM achieved the highest macro recall. A classifier benchmark identified stochastic-gradient-descent logistic regression and ridge-based classification as efficient linear models for the ten-class prediction task from foundation-model-derived embeddings. The findings obtained further informed the development of a sub-feature multi-model ensemble combining Virchow2, H-optimus-1, and GenBio-PathFM. To potentially increase the robustness of the final ensemble, a custom training-set diversification strategy was used. On the hidden validation set, the proposed multi-model ensemble configuration achieved an MCC of 0.893, a F1 score of 0.722, a recall of 0.729, and an accuracy of 0.923, outperforming the challenge baseline strategy.

Keywords: Glioblastoma · Foundation models · Machine learning · Ensembling · Multi-class classification

1 Introduction

The neuropathology of glioblastoma is characterised by spatial heterogeneity [5]. Different regions within a single specimen may show cellular tumour occurring alongside pseudopalisading or geographic necrosis, microvascular proliferation, cortical or white-matter infiltration, leptomeningeal infiltration, and macrophage- or lymphocyte-rich regions. The BraTS-Path challenge [1] defines their recognition as a ten-class histology-patch classification task, including a category for patches outside the nine specified regions of interest. One approach to this task is to represent each histology tile using a pathology foundation

model and train a machine-learning probe on the resulting embedding space. Although such representations can transfer effectively across many histopathology tasks, classification performance may still depend on the nature of the task itself, the choice of the foundation model and the downstream classifier, particularly when class frequencies are imbalanced across sampled image regions and patients. Prior work addressed this setting using class-balanced boosting models trained on separate sub-feature partitions of the embedding and combined at the prediction level [8]. Building on this idea, we developed a lightweight ensemble based on linear classifiers, multiple pathology foundation models, and a custom procedure for increasing downstream training diversity. Before defining the final ensemble members and combination strategies, we systematically compared foundation-model embedding spaces and machine-learning probes, and evaluated sub-feature partitioning, data augmentation, and probability aggregation under common patient-level validation.

2 Methods

The methodological workflow comprised the dataset characterisation, training-data diversification via custom augmentations, patient-level internal evaluation, and selection of foundation models, classifiers, and ensemble components.

2.1 Exploring the training dataset

The training set includes ten histopathological annotation classes: cellular tumour (CT), pseudopalisading necrosis (PN), microvascular proliferation (MP), geographic necrosis (NC), cortical infiltration (IC), white-matter infiltration (WM), leptomeningeal infiltration (LI), dense macrophage regions (DM), lymphocyte presence (PL), and none of the above (NOTA). Class balance differs across patches, slides, and patients (Figure 1). Patient-level evaluation was therefore used to ensure that patches from the same patient did not occur in both training and validation sets, thereby reducing leakage from correlated within-patient samples. DM, LI, and PL have lower patch-level representation, while DM, LI, PL, and NOTA are represented by fewer slides or patients. NOTA is represented by a large number of patches but by comparatively fewer patients, indicating that patch count alone does not capture the breadth of independent case representation. The patch spectra show within-class variation that may arise from biological heterogeneity, technical effects, or annotation uncertainty. These observed differences in class representation and visual appearance may be relevant when evaluating downstream modelling approaches.

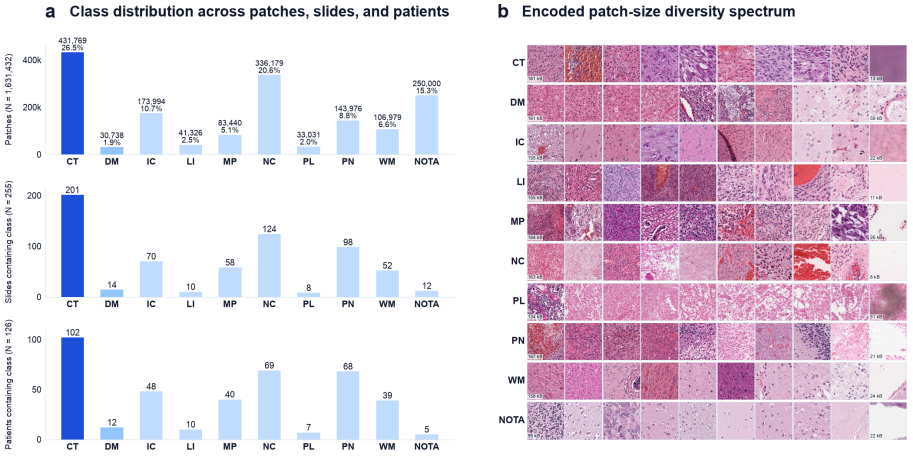

Fig. 1. Training-set composition and patch diversity. (a) Class distributions across patches, slides, and patients. (b) Within-class spectra ordered from largest to smallest encoded patch size, used as a codec-dependent proxy for visual complexity and patch diversity.

2.2 Augmenting the training set to diversify the embedding space

Image augmentation was used to expand the range of representations available during model training. Two augmentation schemes were used. The global scheme targeted patient–class groups with low patch counts. Source patches from these groups underwent stain transfer using donor patches from other patients and classes, followed by rotations or reflections. The number of generated views depended on the representation of the corresponding patient–class group. The local scheme targeted classes with low representation within a given patient and slide. Stain profiles were derived exclusively from the organiser-provided unlabelled training set and summarised at slide level, thus no validation or test data were used to estimate or sample these stain profiles. The global and local schemes generated 140,178 and 147,056 samples, respectively, yielding 287,234 augmented samples per foundation-model embedding set.

Additional training samples were obtained from a publicly available database, the Ivy Glioblastoma Atlas Project (IvyGAP) [9]. Regions corresponding to CT, IC, MP, NC, and PN were sampled randomly and limited to 8,120 image patches per class. IvyGAP augmentation was restricted to classes with sufficiently close histological counterparts in its structural annotation ontology. To our knowledge, the remaining BraTS-Path categories were not found to have sufficiently direct IvyGAP annotations for reliable mapping. These external samples were used only for downstream classifier fitting and were excluded from validation. A limitation of this external augmentation is that IvyGAP supplementation was restricted to five classes and comprised frozen-tissue samples, whereas the challenge data were

derived from formalin-fixed paraffin-embedded tissue. Therefore, class-associated source and tissue preparation effects cannot be excluded.

2.3 Internal evaluation protocol

Candidate methods were evaluated using five-fold patient-level cross-validation. All samples from a given patient were assigned exclusively to either the training or validation set within each fold, thereby preventing overlap between related samples and providing an estimate of performance on previously unseen patients. Class distributions were balanced across folds where permitted by patient-level class support, and an identical fold structure was used for all methods to enable direct comparison. Preprocessing and model fitting were performed independently within each fold. Performance was assessed primarily using Matthews correlation coefficient (MCC), macro-F1, macro recall, weighted one-vs-rest AUROC, and accuracy. Results are reported as mean $\pm$ standard deviation across folds, together with pooled out-of-fold estimates and 95% confidence intervals obtained by patient-cluster bootstrap resampling, with all samples from a patient resampled jointly within their original validation fold. To ensure methodological consistency and computational feasibility across the full set of comparative experiments, this evaluation protocol was applied uniformly throughout.

2.4 Establishing the approach

The modelling approach was established sequentially through foundation-model selection, probe selection, and score-level multi-model ensembling.

Selecting a foundation model for feature extraction Nine pathology foundation models were evaluated as frozen feature extractors: Virchow2 [13], GenBio-PathFM [6], H-optimus-1 [10], UNI2 [2], Prov-GigaPath [12], H0-mini [3], NeuroFM [11], CONCH [7], and Phikon-v2 [4]. Each patch was independently embedded using the corresponding pretrained checkpoint and model-specific preprocessing. The resulting features were L2-normalised and evaluated using identical patient-level folds and a stochastic-gradient-descent logistic regression (SGD–LR) probe. This comparison isolated the effect of the foundation-model representation from downstream classifier choice.

Selecting machine-learning classifiers Following the foundation-model comparison, multiple machine-learning classifiers were evaluated on embeddings from the selected representation space using the same patient-level cross-validation protocol. Candidate probes were assessed according to predictive performance, variation across folds, and computational efficiency. A subset of lightweight and competitive classifiers was selected for the subsequent augmentation and ensemble experiments.

Sub-feature ensembling Embeddings were divided into contiguous blocks of up to 768 dimensions, yielding four blocks for Virchow2 (2560 dimensions; final block 256), two for H-optimus-1 (1536 dimensions), and six for GenBio-PathFM (4608 dimensions). Thus, no embedding dimensions were omitted in the final ensemble. The 768-dimensional width was chosen to limit the dimensionality of individual classifier fits while retaining multiple subspaces for higher-dimensional embeddings. Contiguous blocks provided a deterministic partitioning scheme without introducing an additional random feature-selection step. For each foundation model, an SGD–LR and a Ridge head were fitted to every feature block. SGD–LR contributed class probabilities, whereas Ridge decision-function values were transformed by softmax to normalised class scores. Within each foundation model, block-level scores were equally averaged. The resulting foundation-level vectors were then equally combined and renormalised to obtain the final ensemble-level score.

Decision calibration During model training, a stratified 10% subset of original BraTS-Path training patches was reserved for decision calibration and excluded from classifier fitting. Classes with support at or below the 35th percentile of non-zero class counts received a fixed 1.10 score multiplier, and class-specific thresholds were optimised on the calibration subset by a single coordinate-wise search over 11 values from 0.05 to 2.00 to maximise macro-F1. At inference, ensemble scores were divided by the learned thresholds, renormalised, and the class with the highest adjusted score was assigned. Augmented BraTS-Path and IvyGAP samples were used only for classifier fitting and not for calibration. The same procedure was applied within each cross-validation training fold and when fitting the final model.

Final ensemble training and test-set submission The final ensemble combined sub-feature SGD–LR and Ridge classifiers across Virchow2, H-optimus-1, and GenBio-PathFM, using the described training/calibration partitioning. At inference, feature-space test-time augmentation generated 16 stochastic feature-masking views per patch, each independently retaining 90% of the embedding dimensions, with retained dimensions rescaled by 1/0.9. Ensemble outputs from these views and the unmasked representation were averaged before applying the calibrated decision rule.

3 Results

Foundation-model representations and classifiers were compared first, followed by ablation of the multi-foundation ensemble and evaluation of the selected configuration on the hidden validation set.

3.1 Foundation-model benchmark

The performance of the evaluated foundation-model embedding spaces under the fixed evaluation protocol is reported in Table 1.

Table 1. Foundation-model benchmark using unaugmented embeddings and a fixed SGD log-loss probe.

Model	MCC	Macro-F1	Macro recall	AUROC	Accuracy
Virchow2	**0.715 ± 0.064** **0.718 [0.645 − 0.772]**	0.523 ± 0.067 0.556 [0.490 − 0.583]	0.538 ± 0.081 0.558 [0.503 − 0.601]	**0.948 ± 0.031** **0.942 [0.910 − 0.969]**	**0.774 ± 0.055** **0.775 [0.709 − 0.828]**
GenBio-PathFM	0.705 ± 0.058 0.706 [0.637 − 0.759]	**0.534 ± 0.070** **0.558 [0.487 − 0.587]**	**0.552 ± 0.082** **0.573 [0.526 − 0.607]**	0.942 ± 0.030 0.938 [0.906 − 0.963]	0.765 ± 0.050 0.765 [0.701 − 0.820]
H-optimus-1	0.695 ± 0.054 0.699 [0.632 − 0.753]	0.502 ± 0.060 0.533 [0.465 − 0.571]	0.516 ± 0.050 0.532 [0.491 − 0.573]	0.939 ± 0.034 0.935 [0.900 − 0.965]	0.757 ± 0.050 0.760 [0.697 − 0.816]
UNI2	0.691 ± 0.069 0.692 [0.612 − 0.754]	0.498 ± 0.069 0.513 [0.447 − 0.554]	0.509 ± 0.070 0.529 [0.486 − 0.569]	0.941 ± 0.032 0.935 [0.901 − 0.964]	0.754 ± 0.059 0.754 [0.678 − 0.817]
Prov-GigaPath	0.681 ± 0.083 0.681 [0.605 − 0.746]	0.509 ± 0.071 0.514 [0.448 − 0.571]	0.513 ± 0.062 0.532 [0.490 − 0.574]	0.938 ± 0.036 0.932 [0.897 − 0.961]	0.744 ± 0.076 0.744 [0.667 − 0.813]
H0-mini	0.653 ± 0.071 0.654 [0.580 − 0.715]	0.487 ± 0.077 0.497 [0.426 − 0.540]	0.492 ± 0.081 0.506 [0.466 − 0.555]	0.930 ± 0.040 0.923 [0.883 − 0.955]	0.723 ± 0.061 0.723 [0.652 − 0.787]
NeuroFM	0.652 ± 0.077 0.653 [0.574 − 0.714]	0.476 ± 0.087 0.493 [0.420 − 0.541]	0.488 ± 0.082 0.502 [0.462 − 0.553]	0.923 ± 0.042 0.918 [0.880 − 0.951]	0.724 ± 0.065 0.723 [0.646 − 0.788]
CONCH	0.640 ± 0.061 0.641 [0.569 − 0.696]	0.482 ± 0.068 0.510 [0.437 − 0.545]	0.498 ± 0.074 0.527 [0.474 − 0.565]	0.930 ± 0.033 0.926 [0.890 − 0.952]	0.712 ± 0.049 0.711 [0.642 − 0.768]
Phikon-v2	0.627 ± 0.077 0.629 [0.537 − 0.702]	0.450 ± 0.071 0.474 [0.407 − 0.510]	0.456 ± 0.072 0.482 [0.436 − 0.521]	0.919 ± 0.044 0.912 [0.867 − 0.949]	0.703 ± 0.063 0.703 [0.616 − 0.779]

Virchow2 achieved the highest point estimates for MCC, AUROC, and accuracy, whereas GenBio-PathFM achieved the highest macro-F1 and macro recall. However, the confidence intervals of the leading models overlapped. H-optimus-1 and UNI2 formed a closely matched second tier. The remaining models performed consistently below this leading group.

3.2 Evaluating machine learning classifiers

Linear and non-linear classifiers were first compared using Virchow2 embeddings (Table 2).

Table 2. Classifier comparison using Virchow2 embeddings. SGD–LR used log loss with $\alpha = 3{\times}10^{-5}$, and Ridge used $\alpha = 10$ with the least-squares QR (LSQR) solver. Multilayer perceptron (MLP) used one 256-unit hidden layer, and XGBoost used 50 trees with maximum depth 4.

Classifier	MCC	Macro-F1	Macro recall	AUROC	Accuracy
SGD–LR	**0.715 ± 0.064** **0.718 [0.645 − 0.772]**	0.523 ± 0.067 0.556 [0.490 − 0.583]	0.538 ± 0.081 0.558 [0.503 − 0.601]	**0.948 ± 0.031** **0.942 [0.910 − 0.969]**	**0.774 ± 0.055** **0.775 [0.709 − 0.828]**
Ridge	0.700 ± 0.047 0.705 [0.636 − 0.754]	**0.550 ± 0.039** **0.579 [0.523 − 0.595]**	**0.560 ± 0.059** **0.578 [0.533 − 0.609]**	0.938 ± 0.031 0.936 [0.904 − 0.960]	0.759 ± 0.044 0.762 [0.702 − 0.813]
MLP	0.681 ± 0.082 0.685 [0.618 − 0.743]	0.494 ± 0.080 0.524 [0.451 − 0.565]	0.510 ± 0.077 0.523 [0.483 − 0.581]	0.934 ± 0.036 0.930 [0.899 − 0.956]	0.747 ± 0.073 0.750 [0.682 − 0.813]
XGBoost	0.688 ± 0.070 0.692 [0.617 − 0.748]	0.511 ± 0.070 0.539 [0.474 − 0.570]	0.523 ± 0.074 0.539 [0.486 − 0.580]	0.938 ± 0.033 0.937 [0.908 − 0.961]	0.752 ± 0.062 0.754 [0.687 − 0.811]

SGD–LR achieved the highest point estimates for MCC, AUROC, and accuracy, whereas Ridge achieved the highest macro-F1 and macro recall. The non-linear MLP and XGBoost classifiers did not provide a clear advantage over the linear approaches across the evaluated metrics, suggesting that the foundation-model embedding space was already amenable to relatively simple decision boundaries. Confidence intervals nevertheless overlapped across classifiers, indicating that the observed differences should be interpreted cautiously.

The complementary performance profiles of SGD–LR and Ridge, together with their computational efficiency, motivated their combination in the subsequent ensemble experiments.

3.3 Ensemble design and ablations

To assess the effect of individual design choices on ensemble performance, we performed a stepwise ablation of the main foundation-model combinations, classifier heads, decision calibration, training augmentation, feature partitioning, and test-time augmentation (Table 3). Having established the final ensemble configuration, we next examined its class-specific behaviour using pooled out-of-fold predictions (Table 4). As the final step, the selected ensemble was evaluated on the organiser-provided hidden validation set (Table 5).

Table 3. Ablation of ensemble composition, decision calibration, training augmentation, feature partitioning, and test-time augmentation. VHG denotes the combination of Virchow2, H-optimus-1, and GenBio-PathFM embeddings. SR signifies the combined SGD–LR and Ridge classifier heads. C768-D and R768-D denote contiguous and random 768-dimensional feature blocks, respectively, full-D denotes the full embedding of each foundation model. TTA denotes feature-masking test-time augmentation.

Configuration	Train aug.	MCC	Macro-F1	Macro recall	AUROC	Accuracy
Virchow2–SGD–LR (C768-D, arg max)	—	0.699 ± 0.083 $0.697\,[0.627-0.754]$	0.527 ± 0.080 $0.537\,[0.474-0.585]$	0.554 ± 0.106 $0.559\,[0.511-0.626]$	0.950 ± 0.027 $0.946\,[0.916-0.969]$	0.757 ± 0.076 $0.757\,[0.688-0.817]$
VHG–SGD–LR (C768-D, arg max)	—	0.727 ± 0.061 $0.728\,[0.664-0.778]$	0.547 ± 0.072 $0.570\,[0.505-0.603]$	0.572 ± 0.098 $0.579\,[0.533-0.643]$	0.955 ± 0.025 $0.950\,[0.923-0.974]$	0.782 ± 0.053 $0.782\,[0.722-0.835]$
VHG–SR (C768-D, arg max)	—	0.729 ± 0.057 $0.731\,[0.666-0.780]$	0.553 ± 0.068 $0.581\,[0.515-0.608]$	0.579 ± 0.097 $0.588\,[0.537-0.645]$	0.954 ± 0.027 $0.950\,[0.922-0.972]$	0.785 ± 0.049 $0.784\,[0.725-0.835]$
VHG–SR (C768-D, calibrated)	—	0.733 ± 0.060 $0.735\,[0.666-0.786]$	0.538 ± 0.061 $0.564\,[0.501-0.593]$	0.539 ± 0.066 $0.552\,[0.501-0.588]$	0.947 ± 0.032 $0.941\,[0.909-0.968]$	0.789 ± 0.051 $0.789\,[0.725-0.840]$
VHG–SR (C768-D, calibrated)	Local+Global	0.741 ± 0.059 $0.743\,[0.680-0.793]$	0.547 ± 0.080 $0.578\,[0.519-0.612]$	0.557 ± 0.095 $0.568\,[0.522-0.624]$	0.952 ± 0.030 $0.948\,[0.919-0.972]$	0.795 ± 0.050 $0.795\,[0.736-0.845]$
VHG–SR (C768-D, calibrated)	IvyGAP	0.731 ± 0.060 $0.733\,[0.665-0.785]$	0.536 ± 0.062 $0.562\,[0.501-0.591]$	0.538 ± 0.067 $0.551\,[0.501-0.589]$	0.948 ± 0.028 $0.940\,[0.908-0.968]$	0.787 ± 0.050 $0.787\,[0.725-0.840]$
VHG–SR (C768-D, calibrated)	IvyGAP + Local+Global	0.740 ± 0.058 $0.742\,[0.675-0.793]$	0.546 ± 0.079 $0.576\,[0.515-0.611]$	0.556 ± 0.094 $0.567\,[0.519-0.624]$	0.952 ± 0.029 $0.946\,[0.916-0.971]$	0.794 ± 0.049 $0.794\,[0.733-0.845]$
VHG–SR (R768-D, calibrated)	IvyGAP + Local+Global	0.744 ± 0.058 $0.745\,[0.681-0.795]$	0.550 ± 0.081 $0.579\,[0.526-0.612]$	0.560 ± 0.096 $0.569\,[0.528-0.629]$	0.952 ± 0.030 $0.947\,[0.917-0.973]$	0.797 ± 0.048 $0.797\,[0.737-0.846]$
VHG–SR (full-D, calibrated)	IvyGAP + Local+Global	0.745 ± 0.059 $0.748\,[0.685-0.797]$	0.549 ± 0.075 $0.582\,[0.522-0.614]$	0.560 ± 0.096 $0.572\,[0.522-0.632]$	0.949 ± 0.035 $0.945\,[0.913-0.973]$	0.798 ± 0.051 $0.799\,[0.741-0.848]$
VHG–SR (C768-D, calib.+TTA)	IvyGAP+ Local+Global	0.740 ± 0.058 $0.742\,[0.677-0.792]$	0.545 ± 0.079 $0.576\,[0.517-0.609]$	0.555 ± 0.093 $0.566\,[0.520-0.620]$	0.952 ± 0.030 $0.946\,[0.916-0.972]$	0.795 ± 0.049 $0.794\,[0.735-0.844]$

The largest improvement in internal performance was observed after combining the three foundation-model representations, whereas subsequent components produced smaller changes. Decision calibration slightly increased MCC and accuracy at the expense of macro-F1 and macro recall, while augmentation, alternative feature partitions, IvyGAP supplementation, and feature-masking TTA yielded broadly similar performance with overlapping confidence intervals. Given the relatively marginal differences observed among feature-partitioning strategies, contiguous 768-dimensional blocks were retained as a deterministic and computationally efficient configuration that limited the dimensionality of individual classifier fits. Feature-masking TTA was retained as an inference-time robustness measure, although it produced no material improvement in internal cross-validation.

Table 4. Per-class evaluation of the selected VHG–SR configuration (C768-D, calibrated + TTA, IvyGAP + local + global training augmentation) using pooled out-of-fold predictions. Panel (a) reports the row-normalised confusion matrix as proportions in $[0, 1]$, with shared row labels denoting ground-truth classes and matrix columns denoting predicted classes. Panel (b) reports class-specific precision, recall, and F1. Diagonal confusion-matrix entries are shown in bold.

	(a) Row-normalised confusion matrix										(b) Per-class metrics		
	CT	**DM**	**IC**	**LI**	**MP**	**NC**	**PL**	**PN**	**WM**	**NOTA**	**Prec.**	**Recall**	**F1**
CT	**0.916**	0.000	0.007	0.004	0.018	0.005	0.000	0.028	0.021	0.000	0.772	0.916	0.838
DM	0.149	**0.249**	0.003	0.014	0.031	0.160	0.041	0.242	0.102	0.008	0.681	0.249	0.365
IC	0.054	0.000	**0.860**	0.001	0.003	0.001	0.000	0.003	0.076	0.002	0.877	0.860	0.868
LI	0.783	0.040	0.001	**0.015**	0.062	0.020	0.001	0.077	0.000	0.000	0.131	0.015	0.027
MP	0.388	0.000	0.015	0.014	**0.342**	0.037	0.000	0.158	0.045	0.000	0.625	0.342	0.442
NC	0.002	0.001	0.000	0.000	0.003	**0.941**	0.007	0.046	0.000	0.000	0.906	0.941	0.923
PL	0.226	0.072	0.000	0.201	0.036	0.388	**0.037**	0.039	0.000	0.000	0.010	0.037	0.016
PN	0.110	0.011	0.002	0.001	0.021	0.134	0.000	**0.717**	0.004	0.001	0.660	0.717	0.687
WM	0.199	0.000	0.151	0.001	0.011	0.019	0.000	0.012	**0.589**	0.018	0.678	0.589	0.630
NOTA	0.000	0.000	0.000	0.002	0.000	0.001	0.000	0.000	0.000	**0.996**	0.935	0.996	0.964

Class-specific performance of the selected configuration is reported in Table 4. Observed per-class performance varied across classes. Classification performance was highest for CT, IC, NC, and NOTA, whereas DM, LI, and PL were more difficult. LI was most often misclassified as CT, while PL was most often misclassified as NC or CT.

Table 5. Hidden-validation performance of the proposed multi-foundation ensemble compared with the organiser-provided baseline. VHG–SR denotes the combination of Virchow2, H-optimus-1, and GenBio-PathFM embeddings with sub-feature SGD logistic-regression and Ridge classifiers. Both VHG–SR configurations used identical test-time augmentation settings.

Configuration	**MCC**	**F1**	**Recall**	**AUROC**	**Accuracy**
Organiser baseline	0.689	0.400	0.458	0.871	0.771
VHG–SR + Local Aug.	0.882	0.675	0.662	0.948	0.915
VHG–SR + IvyGAP+Local+Global Aug.	**0.893**	**0.722**	**0.729**	**0.952**	**0.923**

The selected ensemble was evaluated on the organiser-provided hidden validation set (Table 5). Both VHG–SR configurations compared favourably with the organiser-provided baseline across the reported metrics. The configuration including IvyGAP and extended augmentation achieved higher hidden-validation point estimates, especially for macro-F1 and macro recall. Only aggregate metrics were returned for the hidden validation set. Sample-level reference labels were unavailable, which precluded estimation of confidence intervals.

4 Discussion

This study demonstrates that heterogeneous glioblastoma tissue patterns can be classified effectively using frozen pathology foundation-model embeddings and lightweight linear classifiers. This finding is consistent with broader evidence that large self-supervised pathology models provide transferable representations for downstream histology tasks. However, the observed differences among foundation models indicate that representation quality remains task dependent. Virchow2 achieved the highest MCC, AUROC, and accuracy, whereas GenBio-PathFM obtained the best macro-F1 and macro recall, suggesting that models with similar overall performance may encode complementary information relevant to minority or morphologically heterogeneous classes.

SGD logistic regression and Ridge-based classification also showed complementary behaviour. SGD–LR favoured overall agreement and accuracy, whereas Ridge produced higher macro-F1 and macro recall. Combining the two classifiers therefore reduced dependence on a single optimisation objective and was intended to support recognition across classes with unequal representation.

Training-data diversification produced modest changes in internal performance, while IvyGAP supplementation and feature-masking TTA provided little additional improvement under cross-validation. The configuration including IvyGAP achieved higher point estimates on the hidden validation set, although this difference cannot be attributed to IvyGAP alone. External neuropathology data may nevertheless broaden the relevant embedding space, although differences in annotation definitions, tissue processing, and acquisition protocols may also introduce domain mismatch.

Model and ensemble selection relied on the same cross-validation framework, without nested cross-validation, and may therefore have produced optimistic internal estimates. This limitation is partly mitigated by evaluation of the selected system on the organiser-provided hidden validation set, which was independent of the internal cross-validation used for model and ensemble development and therefore provided a separate assessment of potential generalisation. The higher hidden-validation MCC observed may partly reflect training on nearly the full development set, although differences in case composition or difficulty may also have contributed, precluding direct investigation of this discrepancy.

The limited patient-level representation of some classes during training might also reduce confidence in class-specific performance estimates. Despite these constraints, the results show that complementary foundation-model representations, simple linear classifiers, and targeted data diversification might provide an effective and computationally efficient framework for classifying heterogeneous glioblastoma tissue patterns.

Code availability

Code supporting the experiments reported in this study is publicly available at github.com/polejowska/neuropath-fml.

References

1. Bakas, S., Thakur, S.P., Faghani, S., Moassefi, M., Baid, U., Chung, V., Pati, S., Innani, S., Baheti, B., Albrecht, J., Karargyris, A., Kassem, H., Nasrallah, M.P., Ahrendsen, J.T., Barresi, V., Gubbiotti, M.A., López, G.Y., Lucas, C.H.G., Miller, M.L., Cooper, L.A.D., Huse, J.T., Bell, W.R.: BraTS-Path Challenge: Assessing Heterogeneous Histopathologic Brain Tumor Sub-regions (2024)
2. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a General-Purpose Foundation Model for Computational Pathology. Nature Medicine (2024)
3. Filiot, A., Dop, N., Tchita, O., Riou, A., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., Olivier, A.: Distilling foundation models for robust and efficient models in digital pathology (2025), https://arxiv.org/abs/2501.16239
4. Filiot, A., Jacob, P., Kain, A.M., Saillard, C.: Phikon-v2, a large and public feature extractor for biomarker prediction (2024), https://arxiv.org/abs/2409.09173
5. Hambardzumyan, D., Bergers, G.: Glioblastoma: defining tumor niches. Trends in Cancer **1**(4), 252–265 (2015)
6. Kapse, S., Aygün, M., Cole, E., Lundberg, E., Song, L., Xing, E.P.: GenBio-PathFM: A State-of-the-Art Foundation Model for Histopathology. bioRxiv (2026). https://doi.org/10.64898/2026.03.17.712534
7. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
8. Monroy, L.C.R., Mayr, M., Mill, L., Köstler, H., Maier, A.: Patch-Level Brain Tumor Sub-region Classification Using Foundation Models Under Long-Tailed Data Distributions. In: Bakas, S., Dennis, E., Astaraki, M., Baid, U., Conte, G.M., Foltyn-Dumitru, M., Jiang, Z., Labella, D., Metz, M.C., Anazodo, U., de Verdier, M.C., Kofler, F., Li, H.B., Linguraru, M.G., Maleki, N. (eds.) Segmentation, Classification, and Synthesis for Brain Tumors and Traumatic Brain Injuries. pp. 213–222. Springer Nature Switzerland, Cham (2026)
9. Puchalski, R.B., Shah, N., Miller, J., Dalley, R., Nomura, S.R., Yoon, J.G., Smith, K.A., Lankerovich, M., Bertagnolli, D., Bickley, K., Boe, A.F., Brouner, K., Butler, S., Caldejon, S., Chapin, M., Datta, S., Dee, N., Desta, T., Dolbeare, T., Dotson, N., Ebbert, A., Feng, D., Feng, X., Fisher, M., Gee, G., Goldy, J., Gourley, L., Gregor, B.W., Gu, G., Hejazinia, N., Hohmann, J., Hothi, P., Howard, R., Joines, K., Kriedberg, A., Kuan, L., Lau, C., Lee, F., Lee, H., Lemon, T., Long, F., Mastan, N., Mott, E., Murthy, C., Ngo, K., Olson, E., Reding, M., Riley, Z., Rosen, D., Sandman, D., Shapovalova, N., Slaughterbeck, C.R., Sodt, A., Stockdale, G., Szafer, A., Wakeman, W., Wohnoutka, P.E., White, S.J., Marsh, D., Rostomily, R.C., Ng, L., Dang, C., Jones, A., Keogh, B., Gittleman, H.R., Barnholtz-Sloan, J.S., Cimino, P.J., Uppin, M.S., Keene, C.D., Farrokhi, F.R., Lathia, J.D., Berens, M.E., Iavarone, A., Bernard, A., Lein, E., Phillips, J.W., Rostad, S.W., Cobbs, C., Hawrylycz, M.J., Foltz, G.D.: An anatomic transcriptional atlas of human glioblastoma. Science **360**(6389), 660–663 (2018). https://doi.org/10.1126/science.aaf2666, https://www.science.org/doi/abs/10.1126/science.aaf2666
10. Scalbert, M., Saillard, C., Peeters, T., Gonzalez, L., Valter, D., Llinares-López, F., Mariet, Z.E., Jenatton, R.: H-optimus-1: A foundation model for computational histopathology. In: Proceedings of the American Association for Cancer Research Annual Meeting 2026; Part 2 (Late-Breaking, Clinical Trial, and Invited Abstracts). vol. 86, p. LB174. AACR (2026). https://doi.org/10.1158/1538-7445.AM2026-LB174

11. Verma, R., Kandoi, S., Afzal, R., Chen, S., Jegminat, J., Karlovich, M.W., Umphlett, M., Richardson, T.E., Clare, K., Hossain, Q., Samanamud, J., Faust, P.L., Louis, E.D., McKee, A.C., Stein, T.D., Cherry, J.D., Mez, J., McGoldrick, A.C., Mora, D.D.Q., Nirenberg, M.J., Walker, R.H., Mendez, Y., Morgello, S., Dickson, D.W., Murray, M.E., Cordon-Cardo, C., Tsankova, N.M., Walker, J.M., Dangoor, D.K., McQuillan, S., Thorn, E.L., Sanctis, C.D., Li, S., Fuchs, T.J., Farrell, K., Crary, J.F., Campanella, G.: Domain-Specific Foundation Model Improves AI-Based Analysis of Neuropathology (2025), https://arxiv.org/abs/2512.05993
12. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
13. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., Fuchs, T., Fusi, N., Liu, S., Severson, K.: Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology (2024), https://arxiv.org/abs/2408.00738