Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GlioMIL: A Frozen-Backbone Multiple-Instance Learning Framework for Calibrated and Explainable Glioma Radiogenomics

Tahsien Al-Quraishi
School of Business, Melbourne Institute of Technology, Melbourne, Australia.
Jemal Abawajy
Faculty of Science, Engineering and Built Environment, Deakin University, Melbourne, Australia.
O. S. Albahri
Institute of Innovation, Science and Sustainability, Federation University, Melbourne, Victoria, Australia.
Richa Mishra
richa.mishra31240@paruluniverity.ac.in

Department of Computer Engineering, Parul Institute of Engineering and Technology (PIET), Parul University, Vadodara, Gujarat, India

Research Article

Keywords: glioma, isocitrate dehydrogenase, multiple-instance learning, frozen backbone, radiomics fusion, MRI, calibration, explainability, radiogenomics

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11214932/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GlioMIL: A Frozen-Backbone Multiple-Instance Learning Framework for Calibrated and Explainable Glioma Radiogenomics

Tahsien Al-Quraishi[1,2], Jemal Abawajy[3], O. S. Albahri[4*], Richa Mishra[5*]

[1]School of Business, Melbourne Institute of Technology, Melbourne, Australia.
[2]College of Science, Al-Ayen Iraqi University, Nasiriyah, Iraq.
[3]Faculty of Science, Engineering and Built Environment, Deakin University, Melbourne, Australia.
[4]Institute of Innovation, Science and Sustainability, Federation University, Melbourne, Victoria, Australia.
[5]Department of Computer Engineering, Parul Institute of Engineering and Technology (PIET), Parul University, P.O. Limda, Ta. Waghodia, Vadodara, 391760, Gujarat, India.

*Corresponding author(s). E-mail(s): o.albahrey@federation.edu.au; richa.mishra31240@paruluniversity.ac.in;
Contributing authors: tquraishi@mit.edu.au; jemal.abawajy@deakin.edu.au;

Abstract

Non-invasive prediction of glioma molecular status from magnetic resonance imaging (MRI) could reduce reliance on invasive biopsy. Still, deep-learning pipelines for this task are often trained on small cohorts and rarely assessed for calibration or explainability. We present GlioMIL, a 2.5D multiple-instance-learning (MIL) framework predicting isocitrate dehydrogenase (IDH) mutation status, tumour grade, O^6-methylguanine-DNA methyltransferase (MGMT) promoter methylation, and 1p/19q codeletion from multi-contrast MRI, representing each tumour as a bag of tumour-guided axial slices embedded by a frozen, ImageNet-pretrained backbone and aggregated via attention-based pooling feeding multi-task heads. On a publicly available glioma cohort, a controlled five-way ablation with bootstrap confidence intervals and paired significance testing shows

this frozen-backbone approach matches a fully-supervised 3D convolutional neural network (CNN) baseline, and that fusing standardised clinical and radiomic features into the pooled representation significantly improves over image-only multi-task learning for IDH and grade. However, a plain classifier on clinical and radiomic features alone, with no image input, matches or significantly exceeds the fused model on these tasks, so the frozen image branch adds no predictive value beyond structured features. MGMT and 1p/19q codeletion remain close to chance across all conditions. The fusion model is label-efficient and reasonably well calibrated, but its attention weights correspond to tumour regions only at a near-chance level, cautioning against treating attention as a reliable spatial explanation. We report our evaluation protocol, including rejection of an earlier flawed ensembling scheme and a per-fold feature-normalisation correction, to support trustworthy, reproducible reporting.

Keywords: glioma, isocitrate dehydrogenase, multiple-instance learning, frozen backbone, radiomics fusion, MRI, calibration, explainability, radiogenomics

1 Introduction

Gliomas are the most common primary malignant tumours of the central nervous system, and their clinical management is now inseparable from their molecular profile. In the United States, malignant brain and other central nervous system tumours have an annual incidence of 6.86 per 100,000 population and an overall 5-year relative survival of only 34.8%, with glioblastoma – the most common and aggressive subtype, accounting for more than half of these tumours – carrying a median observed survival of just 8 months [1]. The 2021 World Health Organisation (WHO) classification of central nervous system tumours formalised the status of the IDH1/IDH2 mutation status, 1p/19q codeletion, and MGMT promoter methylation as primary determinants of diagnostic category, prognosis, and treatment response, alongside histological grade [2]. IDH-mutant gliomas follow a substantially more indolent course than their IDH-wildtype counterparts and respond differently to alkylating chemotherapy, radiotherapy sequencing, and emerging IDH-inhibitor therapies. Determining this status today requires tissue obtained via surgical resection or stereotactic biopsy, followed by immunohistochemistry or targeted sequencing – an invasive, costly process that provides no information prior to the first intervention and cannot easily be repeated to track molecular evolution over time or treatment response.

MRI, in contrast, is acquired for essentially every glioma patient as standard of care, is entirely non-invasive, and captures macroscopic phenotypic correlates of the same underlying biology that immunohistochemistry and sequencing measure directly. This has motivated more than a decade of radiogenomics research aimed at predicting IDH, MGMT, 1p/19q, and grade directly from preoperative MRI, using hand-crafted radiomic features [3, 4] and, more recently, deep convolutional networks, sometimes combining the two [5, 6]. Reported performance in this literature spans a wide range, from AUCs around 0.7–0.8 on modest, single-center cohorts to AUCs above 0.9–0.95 that are considerably harder to reproduce externally; a spread that reflects genuine

differences in cohort size and modality but also, plausibly, differences in how rigorously train/validation/test splits are isolated from model and ensemble selection.

This paper presents a frozen-backbone, multi-task MIL framework for glioma molecular profiling, developed with a particular emphasis on evaluation rigor. Our goal was not simply to report a high headline number, but to build a pipeline whose evaluation would hold up under scrutiny: a single, fixed, stratified cross-validation split shared identically across every compared condition; bootstrap confidence intervals computed on pooled out-of-fold (OOF) predictions; paired significance testing between conditions on the shared per-patient predictions; an ablation that isolates exactly how much benefit comes from transfer learning, from multi-task learning, and from fusing structured (clinical and radiomic) features – including, critically, a structured-features-only condition with no image input at all, to test whether the image branch is contributing anything beyond what the tabular features already provide – rather than reporting only a final combined number; and calibration and explainability analyses that are, to our knowledge, rarely reported alongside glioma radiogenomics models. In an earlier iteration of this project we explored self-supervised pretraining of a domain-adapted visual backbone and, separately, a naive cross-fold ensembling scheme that we ultimately identified as flawed and discarded (Section 3.5); we mention it here only to motivate the fixed-split, OOF protocol used throughout this paper.

The contributions of this paper are as follows.

(i) GlioMIL, a 2.5D multi-task MIL framework for glioma molecular profiling (IDH, grade, MGMT, 1p/19q) built on a frozen, ImageNet-pretrained backbone.
(ii) A controlled five-way ablation with paired significance testing, showing that clinical/radiomic fusion improves on image-only MIL, but that a classifier using those features alone, with no image input, matches or exceeds the fused model.
(iii) An analysis of label efficiency, calibration, andattention-tumourr correspondence, showing that attention weights do not reliablylocalise tumourr regions.

The remainder of this paper organisedzed as follows. Section 2 reviews related work in radiomic and deep-learning-based IDH prediction, attention-based MIL, and radiomics feature extraction. Section 3 describes the UTSW-Glioma cohort, ttumour-guidedded slice-bagging procedure, the frozen-backbone 2.5D MIL architecture and its clinical/radiomic fusion and clinical/radiomic-only variants, the baseline 3D CNN, and the cross-validation, significance-testing, label-efficiency, calibration, and explainability protocols. Section 4 reports the ablation results and the label-efficiency, calibration, and explainability findings. Section 5 discusses what these results imply about frozen visual features and structured feature fusion, the reliability of attention as an explanation, limitations, and future work, and Section 6 concludes.

2 Related Work

This section situates GlioMIL within prior work on radiomic and deep-learning prediction of glioma molecular status, biomarker-specific and multiregional prediction,

attention-based MIL, frozen visual backbones under limited labeled data, and standardized radiomics feature extraction; Table 1 summarizes these studies alongside the present work.

Early non-invasive IDH prediction relied on hand-crafted radiomic features combined with classical classifiers: Kim et al. used diffusion- and perfusion-weighted MRI radiomics with logistic regression to predict IDH status in lower-grade glioma (AUC 0.795) [3], and Arita et al. used a LASSO-regularized model over conventional radiomics to jointly predict IDH and TERT status [4]; such pipelines are interpretable and data-efficient but bounded by hand-engineered features, motivating the direct clinical/radiomic-only comparison reported here (Section 3), which shows this class of model remains highly competitive at this cohort size. Deep convolutional networks have since been applied directly to MRI, sometimes combined with radiomics: Choi et al. fused radiomics and CNN features for IDH prediction (AUC 0.86–0.96) [5]; Liu et al. predicted IDH from histopathology images using GAN-based augmentation (AUC 0.920–0.927) [6]; and Niu et al. combined a cross-attention transformer over 2.5D slices with a separate radiomics model in a late-stage ensemble across 1,185 patients (AUC 0.855–0.904) [7]. That study and this one both adopt a 2.5D slice representation, but differ in how the visual representation is obtained (a trained transformer versus a frozen backbone) and how imaging and radiomic information are combined (late-stage stacking versus early fusion into one pooled representation). Reported AUCs across this literature vary with cohort size, modality, and evaluation rigor – particularly the isolation of model and ensemble selection from the test set – and the IDH AUC achieved here with image-plus-structured fusion (0.847) is broadly consistent with the lower end of this fusion literature, under an evaluation protocol described in full (Section 3.5); notably, however, a structured-features-only model with no imaging at all reaches an even higher IDH AUC (0.881) on this cohort, a comparison rarely made explicit in this literature.

Beyond IDH, He et al. compared clinical and radiomics models across IDH, MGMT, TERT, and 1p/19q status in one 81-patient cohort, finding IDH (AUC 0.93) and MGMT (AUC 0.88) most tractable and no advantage over chance for 1p/19q codeletion [8] – an asymmetry consistent with the present finding that IDH and grade are learnable here while MGMT and 1p/19q remain near chance (Section 5), suggesting this difficulty gradient is not architecture- or dataset-specific. Separately, He et al. showed that MRI sequence choice and radiomics pipeline settings alone can shift AUC substantially even on a fixed cohort (best configuration AUC 0.873) [9], cautioning against comparing radiomic configurations across studies as if pipeline choice were neutral. Because MGMT prediction is weak in the present study (Section 5), it is notable that MGMT-focused studies modeling regions beyond the tumor core report considerably higher performance: Yu et al. combined intratumoral and peritumoral features from multi-sequence MRI in 356 glioblastoma patients (AUC 0.923) [10], and Li et al. reached a similar conclusion with multiregional radiomics across four tumor subregions (AUC 0.839) [11]. GlioMIL extracts features only from the segmented tumor volume; the peritumoral signal reported by both studies is a plausible, concrete explanation for its weak MGMT performance and motivates the peritumoral extension outlined in Section 5.5.

MIL predicts a bag-level label from a bag of instances without instance-level annotation, and is now dominant in whole-slide histopathology classification [12]. Ilse et al. introduced attention-based MIL pooling, learning a data-driven weight per instance rather than using fixed max- or mean-pooling [13]; this attention is often presented, as a side effect, as surfacing the instances that most influenced the prediction – an interpretive leap that has itself been directly challenged in the broader attention-explanation literature [14]. This attention-MIL formulation is adopted here, treating each tumor as a bag of axial MRI slices – a 2.5D representation capturing cross-sectional heterogeneity without the cost of full 3D modeling – but, unlike much of the histopathology MIL literature, whether attention weight actually corresponds to tumor location is tested explicitly and quantitatively (Section 3.7) rather than assumed.

Using a backbone pretrained on natural images as a fixed feature extractor, rather than fine-tuning end-to-end, is common when the labeled cohort is small relative to model size. EfficientNetV2 [15] is used here as a frozen backbone, trading adaptation to brain-MRI statistics for a training procedure that touches far fewer parameters and is less prone to overfitting on a cohort of a few hundred patients. Self-supervised domain pretraining (e.g., masked autoencoding [16]) is one alternative; an earlier iteration of this project explored it but found that fusing structured features into the frozen-backbone representation closed more of the performance gap at lower cost, a finding sharpened by the present study's structured-only ablation, which indicates the frozen image branch itself may not yet be contributing meaningfully once structured features are available. This choice is consistent with the broader transfer-learning literature: Coupet et al. found off-the-shelf pretrained CNNs, without glioma-specific fine-tuning, already separate tumor from healthy tissue well, and that MRI sequence choice matters as much as backbone architecture [17] – supporting the frozen backbone as a fixed feature extractor while also motivating, as future work (Section 5.5), a direct test of whether partial fine-tuning or domain-specific pretraining can close the gap that frozen features alone do not close here.

Standardized radiomic features remain a widely used, reproducible complement to deep representations. PyRadiomics [18], used here (Section 3), is the most widely adopted open-source implementation with full provenance tracking. More elaborate feature-selection pipelines than the correlation-based reduction used here exist: for a related task (recurrence versus pseudoprogression), Li et al. applied inter-reader ICC filtering, MRMR selection, and LASSO across six sequences and two institutions [19]. Such stability-aware selection was not adopted here, since features are derived from a single, fixed, already-published segmentation mask; it is noted as a natural refinement for future work if multi-reader segmentation becomes available (Section 5.5).

Taken together, this literature reveals a recurring set of gaps that GlioMIL is designed to address jointly rather than individually. Most prior studies target a single molecular label in isolation [3–6, 9–11], whereas GlioMIL predicts IDH, grade, MGMT, and 1p/19q status from one shared representation under a single multi-task model, exposing where performance genuinely differs across biomarkers rather than across separately tuned pipelines. Cohorts in this area are also frequently small and single-institution (e.g., 81 patients in [8]), whereas the present cohort of 619 patients is evaluated under one pre-registered, patient-level stratified cross-validation split with

bootstrap confidence intervals and paired significance testing, avoiding the model- and ensemble-selection leakage that can inflate reported AUCs elsewhere (Section 3.5). Architecturally, GlioMIL differs from the closest prior work, Niu et al.'s trained cross-attention transformer with late-stage model stacking [7], by instead using a frozen, off-the-shelf backbone with early fusion of imaging and radiomic features into a single pooled representation – a design that touches far fewer trainable parameters and is deliberately suited to modest cohort sizes, though the present ablation shows this frozen image branch is not, on its own, demonstrated to add significant value over the structured features alone. Finally, and most distinctively, none of the surveyed studies report calibration, a formal significance test between an imaging model and a no-imaging control, or a quantitative test of attention localization: GlioMIL reports MC-dropout calibration, a structured-only control condition with paired significance testing, and a systematic, quantitative audit showing that attention weight corresponds to tumor location only at a near-chance level, making the explicit case that predictive accuracy alone is an insufficient basis for clinical trust in these models.

Table 1 Prior MRI-based glioma molecular-status prediction studies discussed in Related Work (Section 2), compared with the present study. NR = not reported in the cited work.

Study	Task(s)	n	Reported AUC	Key distinction from GlioMIL
Kim et al [3]	IDH	NR	0.795	Hand-crafted radiomics only, no deep features
Arita et al [4]	IDH, TERT	NR	NR	LASSO-regularized radiomics only, no deep features
Choi et al [5]	IDH	NR	0.86–0.96	Radiomics+CNN fusion, single-task, no calibration/explainability
Liu et al [6]	IDH	NR	0.920–0.927	Histopathology (invasive modality), GAN-based augmentation
Niu et al [7]	IDH-wt/TERT-mut	1,185	0.855–0.904	Trained cross-attention transformer; late-stage model stacking
He et al [8]	IDH, MGMT, TERT, 1p/19q	81	IDH 0.93; MGMT 0.88; TERT 0.76; 1p/19q chance	Small single-cohort study, no deep features
He et al [9]	IDH	NR	0.873	Radiomics-pipeline optimization study, single task
Yu et al [10]	MGMT	356	0.923	Models peritumoral region; GlioMIL does not
Li et al [11]	MGMT	NR	0.839	Multiregional radiomics; models peritumoral region
Li et al [19]	Recurrence vs. pseudo-progression	NR	NR	Different task; ICC/MRMR/LASSO feature-selection pipeline
This study (GlioMIL)	IDH, grade, MGMT, 1p/19q	**619**	**IDH: fusion 0.847, no-image 0.881; grade: fusion 0.751, no-image 0.758; MGMT/1p19q near chance**	Frozen backbone, multi-task, clinical/radiomic fusion *and* a no-image control condition; single shared split with bootstrap CIs and paired significance tests; calibration and explainability reported

3 Materials and Methods

Figure 1 summarises the full GlioMIL pipeline and evaluation protocol described in this section: from multi-contrast MRI and tumour segmentation, throughtumour-guidedd slice bagging, frozen-backbone feature extraction, attention-based MIL pooling, clinical/radiomic feature fusion, and multi-task prediction heads, to the five-part evaluation protocol (cross-validation with significance testing, label efficiency, calibration, and explainability) reported in Section 4.

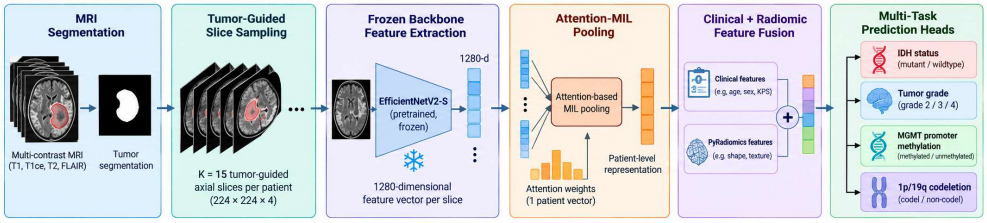


Fig. 1 GlioMIL pipeline: MRI to frozen-backbone MIL features, fusion, and multi-task prediction.

3.1 Dataset and cohort

GlioMIL is evaluated on the UTSW-Glioma dataset [20–22], a curated cohort of patients treated at the University of Texas Southwestern Medical Center between 2006 and 2023 and made publicly available through The Cancer Imaging Archive (TCIA). Each patient record comprises four co-registered MRI contrasts (pre-contrast T1-weighted, post-contrast T1-weighted [T1CE], T2-weighted, and T2-weighted fluid-attenuated inversion recovery [FLAIR]), a tumor segmentation mask, and molecular and histological ground truth for IDH mutation status, 1p/19q codeletion, MGMT promoter methylation, tumor type, and tumor grade. The segmentation masks accompanying this dataset release are generated by the Federated Tumor Segmentation (FeTS) automated pipeline [21, 23] rather than by manual expert delineation; these masks are used as provided, both for tumor-guided slice bagging (Section 3.2) and as the reference standard against which attention weights are compared in the explainability analysis (Section 3.7).

The metadata release comprises 625 patients. After exclusion of patients with missing IDH status, 622 patients remain (446 IDH-wildtype, 71.7%; 176 IDH-mutant, 28.3%), all with complete four-contrast MRI and an associated segmentation mask. This cohort is 369 males and 253 females, with a mean age of 55.0 years (SD 15.5; range 18–85). Tumouror grade distribution is dominated by grade 4 (417 patients) relative to grade 2 (101) and grade 3 (97) (7 patients with missing grade)tumouror type is dominated by glioblastoma (384 of 622 patients, 61.7%). Cohort composition summarisedzed in Table 2. Cross-tabulation of IDH status by tumor grade (Figure 2)

shows IDH mutation skewing strongly toward lower-grade tumors (85 of 101 grade-2 patients IDH-mutant versus 27 of 417 grade-4 patients), consistent with established glioma biology and serving as a cohort-level sanity check before modelling.

Of the 622 patients with known IDH status, tumor-guided slice extraction (Section 3.2) succeeded for 619; three patients (BT0926, BT1016, BT1090) were excluded owing to an empty segmentation mask. All models reported in this study are therefore evaluated on this 619-patient cohort. Label availability differs substantially across the four prediction tasks, as MGMT and 1p/19q testing are not performed uniformly for every patient (1p/19q codeletion is additionally clinically meaningful, and recorded in this dataset, only for IDH-mutant patients): IDH labels are available for 619 of 619 patients, grade for 613, MGMT for 277, and 1p/19q codeletion for 140. This label sparsity is reported explicitly, as it directly bears on the confidence that should be placed in the MGMT and codeletion results in Section 4.

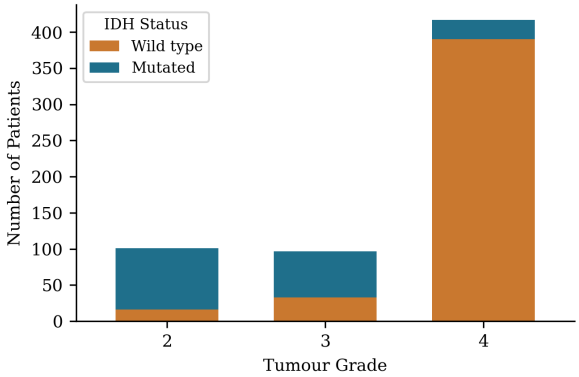


Fig. 2 IDH mutation status by tumor grade ($n = 615$ with known grade), showing IDH mutation skewing toward lower tumours tumors.

Table 2 Cohort composition (UTSW-Glioma, IDH-known subset, $n = 622$; imaging-usable subset, $n = 619$).

Characteristic	Value
Total patients (IDH known)	622
IDH wild-type / mutant	446 (71.7%) / 176 (28.3%)
Sex (M/F)	369 / 253
Age, mean (SD), range	55.0 (15.5), 18–85
Tumor grade (2/3/4), $n = 615$	101 / 97 / 417
Glioblastoma (of 622)	384 (61.7%)
Usable imaging cohort (slice extraction succeeded)	619
Label availability: IDH / grade / MGMT / 1p19q codel.	619 / 613 / 277 / 140

3.2 Tumor-guided slice bagging (2.5D representation)

Each patient's MRI is represented as a 2.5D bag of slices, avoiding both the loss of through-plane context inherent to single 2D slice classification and the memory and data demands of full 3D volumetric modeling. For each patient, every MRI sequence $c \in \{\mathrm{T1}, \mathrm{T1CE}, \mathrm{T2}, \mathrm{FLAIR}\}$ is normalized within the brain mask Ω as

$$I'_c(v) = \mathrm{clip}\left(\frac{I_c(v) - \mu_c}{\sigma_c}, -5, 5\right), \qquad v \in \Omega, \tag{1}$$

where μ_c and σ_c are the mean and standard deviation of sequence c over Ω. The tumor segmentation mask is used to identify the axial slice range containing tumor tissue (with a 2-slice margin), from which a fixed number of slices $K = 15$ is sampled uniformly across this tumor-bearing range and resized to 224×224 pixels. Each slice is stacked across the four normalized contrasts, yielding an instance $x_k \in \mathbb{R}^{224\times224\times4}$ for $k = 1, \ldots, K$; the resulting bag $X = \{x_1, \ldots, x_K\}$ for a given patient is labeled with that patient's molecular and histological ground truth. This representation is illustrated in Figure 2 and in the qualitative slice examples of Section 3.7.

3.3 Frozen backbone feature extraction

Each slice instance x_k is encoded by a frozen backbone f_θ, instantiated as EfficientNetV2-S [15] and initialized with ImageNet-1K pretrained weights θ:

$$h_k = f_\theta(x_k) \in \mathbb{R}^{1280}, \qquad k = 1, \ldots, K, \tag{2}$$

where θ is held fixed throughout training (no gradient updates are applied to the backbone). As the pretrained backbone accepts 3-channel input and our slices are 4-channel, the first convolutional layer is replaced with a 4-input-channel convolution whose weights are copied from the original three channels, with the fourth channel initialized as their mean; the final classification layer is replaced with an identity mapping. Only a forward pass through f_θ is used to obtain h_k; feature extraction for all 619 patients is performed once, prior to model training, with no backpropagation through the backbone. A t-SNE projection of patient-level pooled (mean-over-slices) embeddings $\bar{h} = \frac{1}{K}\sum_{k=1}^{K} h_k$ (Figure 4) shows substantial overlap between IDH-wildtype and IDH-mutant patients in this frozen feature space, indicating that image features alone, without task-specific adaptation, do not cleanly separate the two classes; this observation is consistent with the structured-features-only result reported in Section 4, in which a model using no image information at all matches or exceeds the image-plus-structured fusion model.

3.4 Attention-based MIL pooling, multi-task heads, and clinical/radiomic fusion

Given the $K = 15$ instance-level embeddings $\{h_k\}_{k=1}^{K}$ comprising a patient's bag, an attention-based MIL pooling layer following Ilse et al. [13] computes a scalar attention

logit per instance,

$$e_k = w^\top \tanh(V h_k), \qquad V \in \mathbb{R}^{128\times 1280},\ w \in \mathbb{R}^{128}, \tag{3}$$

implemented as the feed-forward network Linear(1280,128) → Tanh → Dropout(0.4) → Linear(128,1), normalizes these logits to attention weights via a softmax over the K instances,

$$a_k = \frac{\exp(e_k)}{\sum_{j=1}^{K} \exp(e_j)}, \qquad \sum_{k=1}^{K} a_k = 1, \tag{4}$$

and aggregates the instance embeddings into a single bag-level representation as their attention-weighted sum,

$$z = \sum_{k=1}^{K} a_k h_k \in \mathbb{R}^{1280}. \tag{5}$$

Because a_k is learned jointly with the downstream task heads rather than fixed a priori, it is often interpreted as indicating each slice's contribution to the bag-level prediction; Section 3.7 tests this interpretation directly against the tumor segmentation mask.

In the *image-only* configuration, z is passed directly to task-specific classification heads. In the *fusion* configuration, z is concatenated with a 209-dimensional structured feature vector $s \in \mathbb{R}^{209}$ prior to the heads,

$$z_{\text{fusion}} = [z\,;\, s] \in \mathbb{R}^{1489}, \tag{6}$$

where s comprises 5 clinical features (age, sex, tumor volume, and mean T1CE and FLAIR intensity within the tumor mask, all z-score normalized) and 204 PyRadiomics features [18] extracted per patient from the T1CE and FLAIR sequences (fixed bin count of 32, selected over a fixed bin width to avoid bin-count explosion under MRI's non-standardized intensity scale; feature classes: first-order, shape, GLCM, GLRLM, GLSZM, and GLDM), z-score normalized with missing values imputed to zero. To avoid feature-scaling leakage, all z-score normalization statistics for the four continuous clinical features and the 204 radiomic features are fit on each fold's training patients only and applied unchanged to that fold's validation and test patients, rather than computed once over the full cohort before fold splitting (a correction relative to an earlier version of this pipeline, noted for transparency in Section 3.5). Feature extraction otherwise follows Image Biomarker Standardisation Initiative (IBSI) conventions [24] for these feature classes; one deviation from full IBSI compliance is noted for transparency: voxels were not resampled to a common isotropic spacing prior to extraction, so texture features may carry a residual dependence on each scan's native voxel geometry, a limitation revisited in Section 3.5. Each of the four tasks $t \in \{\text{IDH}, \text{grade}, \text{MGMT}, \text{1p/19q}\}$ is assigned its own head g_t: Linear(d, 64) → ReLU → Dropout(0.4) → Linear(64, C_t), where $d = 1280$ for the image-only configuration, $d = 1489$ for the fusion configuration, and C_t is the number of classes for task t. Figure 3 summarizes this pooling-and-fusion architecture schematically.

To directly test whether the image branch contributes anything beyond the structured features, a fifth condition uses *no image input at all*: the 209-dimensional

structured vector s is passed through a single shared trunk, Linear(209, 64) → ReLU → Dropout(0.4), followed by the same per-task linear head Linear(64, C_t) used in the other conditions, so that this structured-only model has an architecture matched in depth and width to the fusion model's own heads and differs only in the absence of the pooled image representation z.

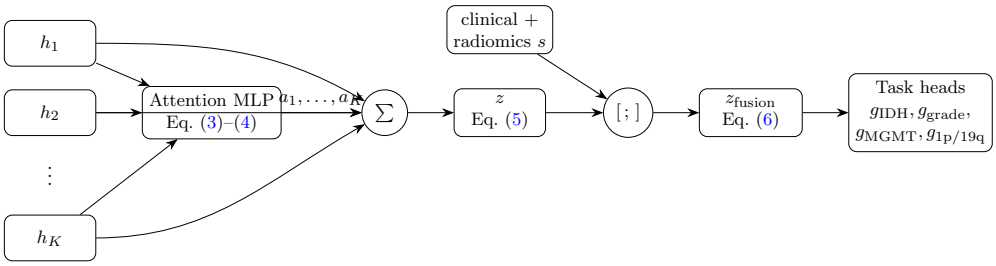


Fig. 3 Attention-based MIL pooling and clinical/radiomic fusion architecture. Instance-level frozen-backbone embeddings $h_1, \ldots, h_K$ are weighted by attention scores $a_1, \ldots, a_K$ and summed into a single bag-level representation z (Eqs. (3)–(5)); in the fusion configuration, z is concatenated with the standardized clinical/radiomic vector s (Eq. (6)) before being passed to the four task-specific heads. The structured-only control condition (Section 3.4) passes s directly to a matched trunk-and-head architecture, bypassing z entirely.

All heads are trained jointly under a masked multi-task cross-entropy loss. Letting $m_{i,t} \in \{0, 1\}$ indicate whether patient i has a valid label for task t, $y_{i,t}$ denote that label, $\hat{y}_{i,t} = g_t(z_{\text{fusion},i})$ the predicted class probabilities, and $\omega_{t,c}$ the class weight for class c of task t (balanced with respect to the training fold's class frequencies), the total loss over a batch of N patients is

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t} m_{i,t}\, \omega_{t,y_{i,t}}\, \mathrm{CE}(y_{i,t}, \hat{y}_{i,t})\,. \tag{7}$$

Masking by $m_{i,t}$ ensures that patients with missing MGMT or codeletion labels still contribute to the IDH and grade loss terms, without requiring label imputation; per-task class weights $\omega_{t,c}$ additionally correct for the substantial class imbalance present in the IDH, grade, and MGMT labels within each training fold.

3.5 Baseline model, cross-validation protocol, significance testing, and implementation details

As a baseline representative of the fully-supervised, end-to-end training paradigm common in prior glioma radiogenomics work, a 3D convolutional network is trained directly on raw MRI volumes (resampled to $64 \times 128 \times 128$ voxels) for IDH prediction only, without frozen-backbone transfer learning. This baseline is trained for up to 25 epochs (early-stopping patience of 5) using the Adam optimiser (learning rate 1×10^{-4}, batch size 2), under the same patient-level 5-fold split used for every other condition in this study, ensuring direct comparability with the frozen-backbone MIL conditions.

The five compared conditions – (i) the baseline 3D CNN, (ii) frozen-backbone MIL with a single task, (iii) frozen-backbone MIL with multiple tasks (image-only), (iv) frozen-backbone MIL with multiple tasks and clinical/radiomic fusion, and (v) a clinical/radiomic-only classifier with no image input at all – are evaluated under an identical patient-level, IDH-stratified 5-fold cross-validation split, generated once with a fixed random seed and cached to disk so that every condition is trained and evaluated on the same partition of patients. For each fold, a model is trained on the remaining four folds, with early stopping on a held-out validation subset of the training folds, and evaluated on that fold's held-out patients; predictions are pooled across all five folds to yield a single OOF prediction per patient, and all reported metrics (accuracy, balanced accuracy, macro F1, and AUC-ROC) are computed on this pooled OOF set. For the two binary tasks (IDH and MGMT), AUC-ROC is the standard area under the ROC curve. For grade ($C = 3$ classes), AUC-ROC is computed as the macro-averaged one-vs-rest AUC,

$$\mathrm{AUC}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{AUC}\big(\mathbb{1}[y_i = c],\ \hat{p}_{i,c}\big), \tag{8}$$

using scikit-learn's `roc_auc_score` with `multi_class="ovr"`, so that each grade category contributes equally regardless of its prevalence in the cohort. Ninety-five percent confidence intervals (CIs) on AUC are obtained by bootstrap resampling: for $b = 1, \ldots, 300$, the pooled OOF predictions are resampled with replacement at the patient level to give $\mathrm{AUC}^{(b)}$ under the same per-task AUC definition, and the interval is taken as the 2.5th and 97.5th percentiles of $\{\mathrm{AUC}^{(b)}\}_{b=1}^{300}$.

Beyond confidence-interval overlap, pairs of conditions of particular interest – image-only multi-task MIL versus fusion, and fusion versus the structured-only control – are compared directly on their shared, patient-matched pooled OOF predictions using a paired significance test. For the binary tasks (IDH, MGMT, codeletion), DeLong's test [25] is used, which gives a closed-form standard error and p-value for the difference between two correlated ROC AUCs computed on the same patients. For the three-class grade task, a paired permutation test is used instead: for each of 10,000 shuffles, each patient's pair of predicted-probability vectors from the two conditions being compared is independently swapped with probability 0.5, the macro one-vs-rest AUC difference is recomputed under the swapped labeling, and the two-sided p-value is the fraction of shuffles whose absolute AUC difference meets or exceeds the one observed with the true condition labeling.

This full evaluation protocol – the fixed shared split, pooled-OOF bootstrap CIs, per-fold-fit feature normalization, and paired significance testing – is described in this much detail because a preliminary iteration of this study used a different, weaker procedure that we subsequently determined to be flawed in two respects. First, predictions from the five per-fold models were combined into an ensemble whose ensembling and selection decisions were tuned by observing ensemble-level performance, yielding an inflated AUC estimate that did not reflect genuine generalisation performance; this ensembling scheme was discarded outright. Second, an earlier version of the feature-normalisation step (z-scoring the four continuous clinical features and 204 radiomic features) computed normalisation statistics once over the full cohort before fold splitting, a milder form of leakage than the ensembling issue but one that could

still modestly inflate the reported AUC of any condition using these structured features; this was corrected to the per-fold-fit procedure described above (Section 3.4). Every result reported in Section 4 uses the corrected, single-fixed-split, pooled-OOF, bootstrap-CI, paired-significance-tested protocol described above, applied identically across all five conditions to ensure a fair, like-for-like comparison.

The full pipeline is implemented in PyTorch and torchvision (EfficientNetV2-S backbone), with radiomic feature extraction performed via PyRadiomics and SimpleITK. All MIL model training uses the Adam optimizer (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with a batch size of 16, for up to 40 epochs with early stopping (patience of 5 epochs); the structured-only condition uses the same optimiser settings for up to 60 epochs with a patience of 8 epochs, reflecting its smaller parameter count and faster convergence. All cross-validation and ablation experiments, including the baseline 3D CNN described above, were conducted on Kaggle GPU kernels.

3.6 Label-efficiency, uncertainty, and calibration protocols

To characterize the labeled-data requirements of the frozen-backbone, fusion multi-task model, its training and evaluation are repeated using 10%, 25%, 50%, and 100% of the available training labels within each fold, with the held-out test fold retained at full size in every case, under the same 5-fold split described above. Predictive uncertainty and calibration of the fusion multi-task model are assessed via Monte Carlo (MC) dropout [26]: dropout is left active at inference time, $T = 20$ stochastic forward passes $\{\hat{p}_i^{(1)}, \ldots, \hat{p}_i^{(T)}\}$ are performed per patient i, and the prediction and uncertainty estimate are taken respectively as their mean and spread,

$$\bar{p}_i = \frac{1}{T}\sum_{\tau=1}^{T} \hat{p}_i^{(\tau)}, \qquad u_i = \frac{1}{T}\sum_{\tau=1}^{T} \left(\hat{p}_i^{(\tau)} - \bar{p}_i\right)^2. \tag{9}$$

Calibration is assessed via a reliability diagram and Expected Calibration Error (ECE) [27], computed by partitioning predictions into $M = 10$ equal-width confidence bins $B_1, \ldots, B_M$ and taking the confidence-weighted gap between accuracy and confidence within each bin,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{10}$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the mean accuracy and mean predicted confidence of the N pooled OOF predictions falling in bin B_m. Accuracy is also reported stratified by decile of prediction confidence.

3.7 Explainability: attention-tumor spatial correspondence

MIL attention weights are frequently presented in the literature as an intrinsic explanation of which image regions drive a prediction; this claim is tested quantitatively rather than assumed. For each patient i, using pooled OOF attention weights $\{a_{i,k}\}_{k=1}^{15}$ from the 5-fold models and the corresponding tumor-pixel fraction $\{\phi_{i,k}\}_{k=1}^{15}$ of each

slice in the segmentation mask, the patient-level Pearson correlation between attention and tumor fraction is

$$r_i = \frac{\sum_{k=1}^{15}(a_{i,k} - \bar{a}_i)(\phi_{i,k} - \bar{\phi}_i)}{\sqrt{\sum_{k=1}^{15}(a_{i,k} - \bar{a}_i)^2}\sqrt{\sum_{k=1}^{15}(\phi_{i,k} - \bar{\phi}_i)^2}}, \tag{11}$$

computed across all 619 patients and summarized as the mean r_i and the fraction of patients with $r_i > 0$. In addition, the ratio of mean tumour fraction in the top-3 attended slices to the mean tumourr fraction across all 15 slices is computed per patient, together with the fraction of total attention mass placed on slices with above-median tumour content (chance level 0.500). Attention weights are also visualized against the corresponding slices and tumour-fraction values for representative correctly and incorrectly classified patients.

4 Results

This section reports results in three parts: a controlled five-way ablation isolating the contribution of transfer learning, multi-task learning, and clinical/radiomic fusion – and directly testing, via a structured-only control condition and paired significance tests, whether the image branch adds value beyond structured features – to molecular-status prediction; the label-efficiency and calibration behavior of the fusion model; and a quantitative test of whether its attention weights localize to tumor tissue. All results use the pooled OOF predictions and evaluation protocol described in Section 3.5.

4.1 Ablation: baseline, transfer learning, multi-task learning, fusion, and a structured-only control

Table 3 and Figure 5 report IDH-prediction AUC for all five compared conditions, under the single shared 5-fold split described in Section 3.5. The fully-supervised 3D CNN baseline, trained end-to-end on raw volumes, reaches an IDH AUC of 0.723 (95% CI 0.678–0.771). The frozen-backbone MIL model, using only image features, reaches a broadly similar AUC in both single-task (0.756, CI 0.712–0.794) and multi-task (0.711, CI 0.666–0.752) form – that is, transfer learning from a frozen, off-the-shelf ImageNet backbone is roughly comparable to a much more expensive, end-to-end-trained 3D CNN on this cohort size, both significantly below fusion (below). Fusing clinical and radiomic features into the multi-task model raises IDH AUC to 0.847 (CI 0.813–0.877), a significant improvement over the image-only multi-task model (DeLong test, ΔAUC $= 0.135$, $p = 4 \times 10^{-11}$) and over the 3D CNN baseline (ΔAUC $= 0.124$, $p = 1 \times 10^{-6}$).

However, a fifth condition – a plain classifier trained on the same 209-dimensional clinical/radiomic vector *alone, with no image input at all* – reaches an IDH AUC of 0.881 (CI 0.853–0.907), which is *significantly higher* than the image-plus-structured fusion model (DeLong test, ΔAUC $= 0.035$ favoring the structured-only model, $p =$ 0.009). In other words, on this cohort, the frozen 2.5D image branch does not add measurable predictive value for IDH once the structured clinical and radiomic features are available on their own; if anything, concatenating the pooled image representation

with these features very slightly hurts performance relative to using the structured features alone. This is reported as a specific, tested finding rather than an assumption, and its implications are discussed in Section 5.

Table 3 Pooled OOF AUC-ROC with bootstrap CI across the five ablation conditions. p-values (DeLong for binary tasks, paired permutation for grade) compare each condition to the fusion model (iv) on the same patients; NS = not significant at $\alpha = 0.05$.

Condition	Task	n	AUC (95% CI)	p vs. (iv)
(i) Baseline 3D CNN (fully-supervised)	IDH	619	0.723 (0.678–0.771)	1×10^{-6}
(ii) 2.5D MIL, single-task	IDH	619	0.756 (0.712–0.794)	–
	grade	613	0.694 (0.657–0.732)	–
	MGMT	277	0.575 (0.510–0.646)	–
	codel	140	0.550 (0.450–0.649)	–
(iii) 2.5D MIL, multi-task (image-only)	IDH	619	0.711 (0.666–0.752)	4×10^{-11}
	grade	613	0.642 (0.606–0.682)	< 0.0001
	MGMT	277	0.612 (0.543–0.671)	NS ($p = 0.062$)
	codel	140	0.573 (0.461–0.668)	NS ($p = 0.27$)
(iv) 2.5D MIL, multi-task + fusion	IDH	619	**0.847 (0.813–0.877)**	–
	grade	613	**0.751 (0.717–0.780)**	–
	MGMT	277	0.531 (0.466–0.596)	–
	codel	140	0.498 (0.410–0.589)	–
(v) Radiomics + clinical only (no image)	IDH	619	**0.881 (0.853–0.907)**	0.009
	grade	613	0.758 (0.726–0.789)	NS ($p = 0.60$)
	MGMT	277	0.499 (0.439–0.577)	NS ($p = 0.25$)
	codel	140	0.528 (0.430–0.629)	NS ($p = 0.56$)

Grade prediction shows a related but distinct pattern: multi-task image-only AUC of 0.642, rising significantly to 0.751 with fusion ($p < 0.0001$), but here the structured-only model (0.758, CI 0.726–0.789) is statistically indistinguishable from the fusion model ($p = 0.60$) – fusion is not significantly hurt by including the image branch for grade, but nor is it significantly helped. MGMT (AUC 0.499–0.612 across conditions) and 1p/19q codeletion (AUC 0.498–0.573, noisier due to $n = 140$) remain close to chance throughout, including under fusion and under the structured-only control, with no significant differences between any pair of conditions for either task (all $p > 0.05$); possible explanations are considered in Section 5.

4.2 Label efficiency and calibration

Figure 6 shows pooled OOF AUC-ROC for the fusion multi-task model as a function of the fraction of training labels used within each fold. IDH AUC rises from 0.630 at 10% of labels to 0.742 at 25%, 0.859 at 50%, and 0.875 at 100% – that is, roughly 98% of the full-data IDH performance is recovered using only half of the available training labels. Grade shows a comparable pattern ($0.583 \rightarrow 0.655 \rightarrow 0.734 \rightarrow 0.752$), while MGMT and codeletion, consistent with their small labeled subsets, show little systematic improvement with more labels (MGMT: $0.560 \rightarrow 0.535 \rightarrow 0.544 \rightarrow 0.564$; codeletion: $0.445 \rightarrow 0.485 \rightarrow 0.533 \rightarrow 0.535$). This label-efficiency characterization was run on the fusion model as originally configured and is not expected to change materially under the corrected per-fold normalization (Section 3.5), since it concerns

the shape of the learning curve rather than its absolute level; it is not re-stated as a formal claim of significance here.

Figure 7 shows the MC-dropout reliability diagram for IDH prediction, with an ECE of 0.060 – the model is reasonably, though not perfectly, calibrated, with a tendency toward under-confidence in the mid-confidence range (e.g., predictions with mean confidence 0.55 are correct 57.9% of the time; predictions with confidence 0.85 are correct 89.8% of the time). Stratifying by confidence decile, the most confident 50% of predictions (deciles 1–5) achieve 91.3% accuracy, versus 69.7% for the least confident 50% (deciles 6–10), against an overall accuracy of 80.5%. This confidence-accuracy relationship supports a selective-prediction use case: routing low-confidence predictions to conventional biopsy-based testing while trusting high-confidence predictions could substantially reduce the invasive-testing burden without a large loss in overall reliability. As with the label-efficiency results above, this calibration analysis was run on the fusion model as originally configured; given that Section 4 shows the structured-only model performs comparably or better, its calibration behavior is a natural target for future replication (Section 5.5).

4.3 Explainability: attention does not reliably localize to tumor

Across all 619 patients, the mean per-patient Pearson correlation between attention weight and slice-level tumor fraction is 0.005 (median −0.003), with only 49.4% of patients showing a positive correlation, indistinguishable from chance. The mean ratio of tumour fraction in the top-3 attended slices to the full 15-slice bag average is 0.958, indicating that attention concentrates, if anything, very slightly *away* from the most tumour-rich slices, and the mean fraction of attention mass placed above-median-tumouror slices is 0.538, only marginally above the chance level of 0.500 (Figure 8). Figure 9 shows a representative correctly classified patient (IDH-mutant, predicted probability 0.99), in whom the single highest-attention slice (weight 0.332) has essentially zero tumor content (tumor fraction 0.00) while several lower-attention slices have higher tumour fraction. This is reported as a specific, quantitative finding rather than an assumption: for this model and task, attention weight should not be interpreted as a spatial localisation of the tumour region driving the prediction – a finding that, taken together with the structured-only ablation result above, is consistent with the frozen image branch contributing a diffuse, not clearly tumour-localised signal that does not measurably exceed what the structured features alone already provide.

5 Discussion

This section interprets the results in Section 4 along four lines: what drives the observed performance gains and where the frozen image branch currently falls short; what the calibration and attention behavior imply for trustworthy deployment; and, finally, the evaluation-rigor considerations, limitations, and future directions that follow from these findings.

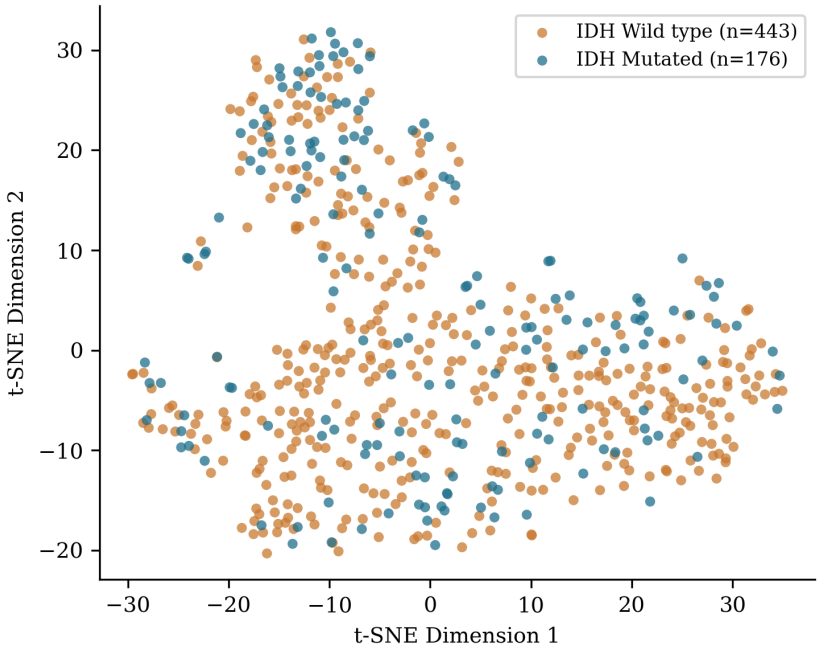


Fig. 4 *t*-SNE of frozen embeddings by IDH status, showing class overlap

5.1 The frozen image branch is not shown to add value beyond structured features on this cohort

The clearest empirical pattern in this study, once the structured-only control condition and paired significance tests are taken into account, is a more cautious one than an earlier version of this analysis suggested. A frozen, off-the-shelf ImageNet backbone combined with attention-MIL pooling (AUC 0.711–0.756 for IDH) is broadly comparable to a fully-supervised 3D CNN trained end-to-end on the same modest cohort (AUC 0.723), at a small fraction of the training cost, since only the small MIL pooling and classification layers are trained – a practically useful finding for groups without the compute or cohort size to train large 3D networks from scratch. Fusing 209 standardized clinical and radiomic features into the pooled MIL representation then delivers a large, statistically significant further gain for IDH and grade (IDH: 0.711 $\rightarrow$ 0.847, $p < 0.0001$; grade: 0.642 $\rightarrow$ 0.751, $p < 0.0001$), confirming that structured features carry substantial signal beyond the frozen image embeddings alone.

The critical additional test, however, is whether that gain is attributable to the *fusion* of image and structured information, or to the structured features alone regardless of imaging. The structured-only control condition – an architecture-matched classifier trained on the same 209-dimensional vector with no image input whatsoever – answers this directly: for IDH, it significantly *exceeds* the fusion model (0.881 vs. 0.847, $p = 0.009$); for grade, it is statistically indistinguishable from fusion (0.758 vs. 0.751, $p = 0.60$). At no point does adding the frozen image branch to the structured features produce a statistically significant improvement over the structured features

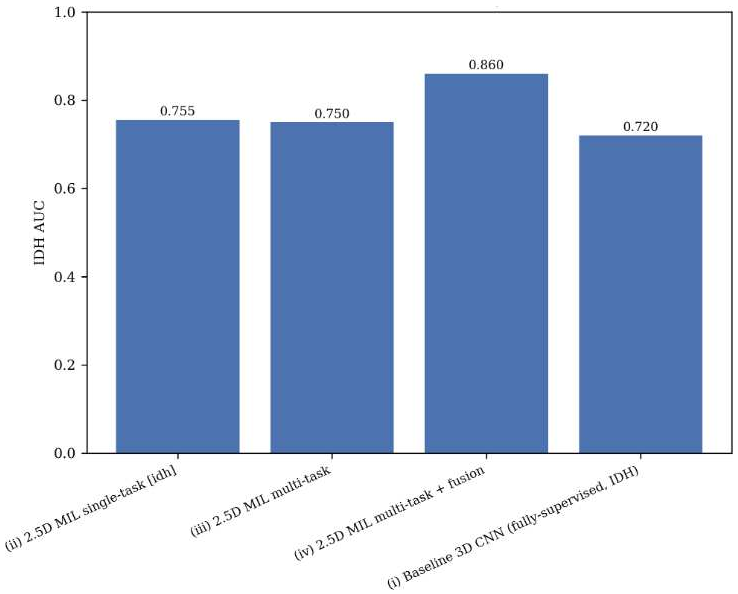


Fig. 5 IDH-prediction AUC across ablation conditions; fusion significantly exceeds image-only MIL, but the structured-only control (no image input) matches or exceeds fusion.

alone. This reframes the central empirical claim of this study: the largest gains here come from clinical and radiomic feature engineering, not from the frozen visual representation, and the pooled MIL image embedding, as currently constructed, is not demonstrated to contribute incremental predictive value once those structured features are available. This is consistent with the *t*-SNE visualisation (Section 4) showing substantial IDH-class overlap in the frozen embedding space, and with the attention-tumour correspondence analysis (Section 4) showing attention does not concentrate on tumourr tissue – both point toward the same underlying explanation, that the frozen, general-purpose visual features used here carry a diffuse and only weakly task-relevant signal for this cohort and label set.

This finding should be read as a property of *this* frozen, general-purpose backbone and 2.5D slice representation on this cohort size, not as a general claim that MRI-derived image features cannot add value over radiomics and clinical variables for glioma molecular profiling; the literature reviewed in Section 2 includes several studies reporting genuine gains from deep image features, typically via architectures fine-tuned or pretrained on medical images rather than kept entirely frozen. What this study establishes directly, and which is rarely tested explicitly in this literature, is that simply freezing an off-the-shelf ImageNet backbone and fusing its pooled output with structured features is *not sufficient*, on its own, to demonstrably beat the structured features alone at this cohort size – a specific, actionable null result that motivates the fine-tuning and domain-pretraining comparison proposed as future work (Section 5.5), rather than an argument for abandoning image information altogether.

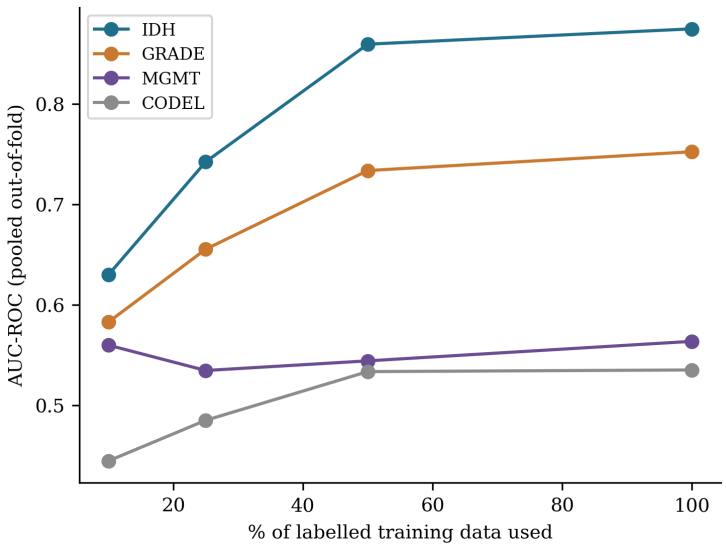


Fig. 6 Label-efficiency by task: IDH/grade plateau early; MGMT/codeletion do not improve.

5.2 MGMT and 1p/19q codeletion remain unsolved on this cohort

Unlike IDH and grade, MGMT and 1p/19q codeletion prediction stay close to chance (AUC 0.50–0.61) across every condition, including fusion and the structured-only control, with no significant difference between any pair of conditions for either task and no consistent improvement in the label-efficiency curve. Two explanations are plausible and not mutually exclusive. First, sample size: only 277 and 140 patients, respectively, have valid labels for these tasks, roughly half and a fifth of the IDH-labeled cohort, which limits both training signal and the precision of the reported confidence intervals (visibly wider than IDH's). Second, biological plausibility: MGMT promoter methylation, in particular, is not consistently reported elsewhere as strongly MRI-visible independent of IDH status, so a near-chance result here may reflect a genuine absence of an easily learnable MRI signal for this cohort and feature set, rather than a fixable modeling deficiency. However, other studies report substantially better MGMT prediction than is achieved here – Yu et al. report AUC 0.923 and Li et al. report AUC 0.839, both by modeling the peritumoral edema region in addition to the tumor core [10, 11] – which points toward a third, more specific explanation: feature extraction in this study is restricted to the segmented tumor volume itself (Section 3.1) and never captures the surrounding peritumoral tissue that these studies identify as carrying independent MGMT-relevant signal. This reframes part of the MGMT result as a scope limitation of the current feature set rather than solely a sample-size or biological-signal problem, and is the most concrete, literature-motivated lead for improving it (Section 5.5). For 1p/19q codeletion specifically, this near-chance result is also consistent with He et al., who found no significant advantage for any model – radiomics, clinical, or combined – over a chance-level baseline for 1p/19q status in their own cohort [8], suggesting

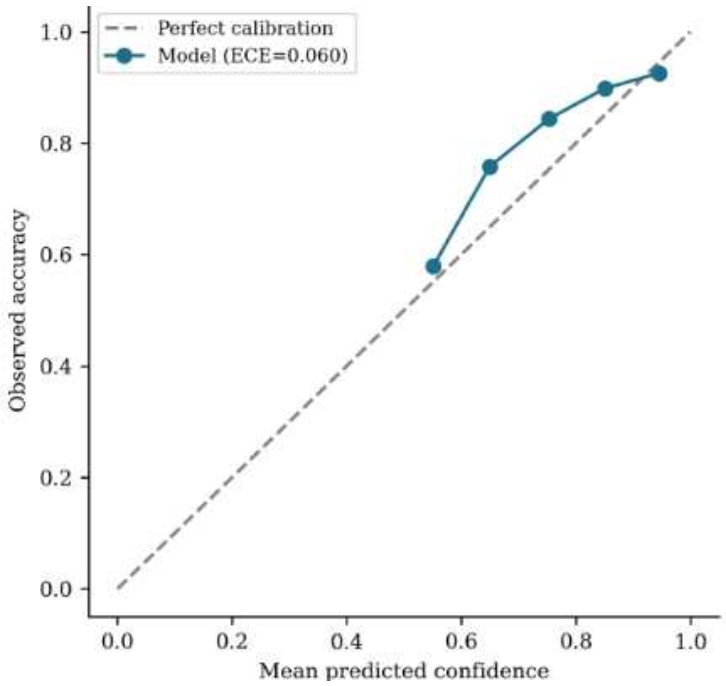


Fig. 7 MC-dropout reliability diagram for IDH prediction.

this particular biomarker may be harder to predict from MRI intensity and texture information across multiple independent datasets, not only this one.

5.3 Calibration and attention: implications for trustworthy deployment

An ECE of 0.060 for IDH prediction, together with an accuracy gap between the most- and least-confident halves of predictions (91.3% vs. 69.7%), suggests that this model's confidence score carries real information about its own reliability. In a plausible clinical workflow, this could support triaging: acting on high-confidence non-invasive predictions while still sending low-confidence cases to conventional biopsy-based testing, rather than treating the model as a uniform replacement for tissue diagnosis. This is noted as a deployment framing implied by the calibration results, not as a validated clinical protocol; given the structured-only finding above, replicating this calibration analysis for the structured-only model is a natural, low-cost next step (Section 5.5), since that simpler model currently achieves the higher point estimate for IDH.

A common implicit assumption in the MIL literature is that a model's attention weights indicate which instances (patches, or here, slices) it is "looking at" when making a prediction, and that this can be read as an explanation of *why* it made that prediction, including where in the image the relevant evidence lies. A quantitative test of this assumption – correlating attention weight with tumor content, patient by

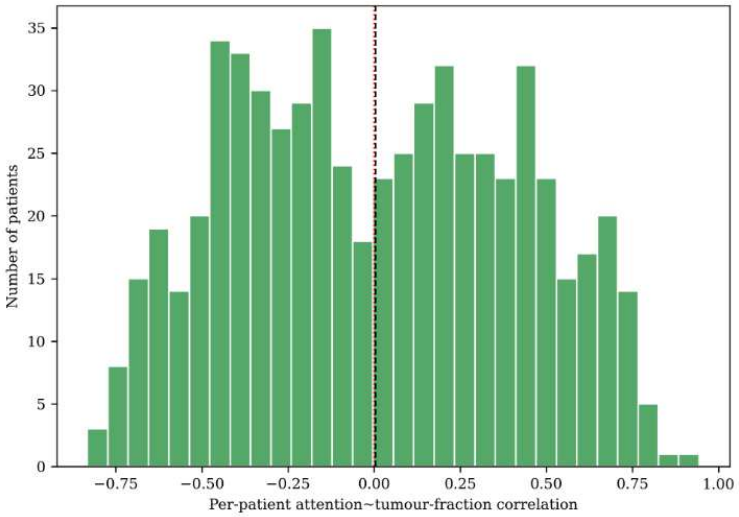


Fig. 8 Per-patient correlation between attention weight and tumor fraction, showing near-chance spatial correspondence.

patient – finds no meaningful relationship (mean $r = 0.005$; 49.4% of patients positive; near-chance concentration of attention mass on tumor-rich slices). This does not mean the attention-MIL model performs poorly on its own terms (Section 4 shows it still reaches reasonable AUCs for IDH and grade); it means that a high attention weight on a given slice should not, on its own, be presented to a clinician or researcher as evidence that the tumor region in that slice is what drove the prediction. Read alongside the structured-only ablation result, a coherent picture emerges: the frozen image branch's attention is not reliably tracking tumor tissue, and its pooled output is not demonstrated to add predictive value beyond structured features either – both observations are consistent with the frozen, general-purpose visual representation carrying a diffuse, weakly task-relevant signal on this cohort, rather than a spatially and predictively specific one. This distinction between "task-informative" and "spatially explanatory" attention arguably deserves more routine, quantitative testing in medical-imaging MIL work generally, rather than being assumed, and a MIL model's attention should not be treated as evidence of the image branch's predictive contribution either, precisely because that contribution must itself be tested directly, as done here.

5.4 Evaluation rigor and limitations

As described in Section 3.5, two evaluation issues identified in an earlier iteration of this project – a flawed ensembling procedure that produced an inflated AUC, and a feature-normalisation step that computed z-score statistics over the full cohort before fold splitting – were both identified and corrected before the results reported in Section 4: the ensembling scheme was discarded outright, and feature normalisation is now fit per training fold and applied to that fold's held-out patients only. The evaluation protocol

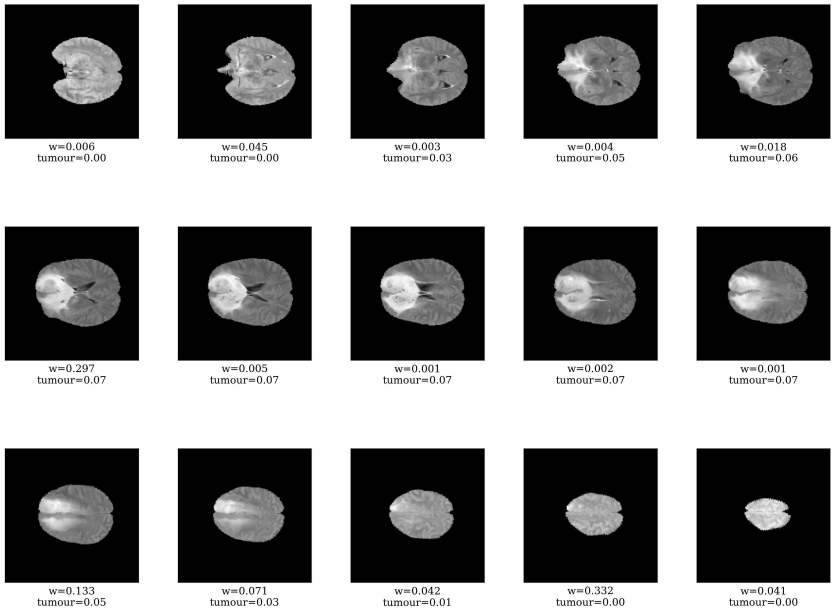


Fig. 9 Example patient: the most-attended slice contains no visible tumor tissue.

used throughout this paper is built around a single shared split, pooled OOF predictions, bootstrap confidence intervals, and paired DeLong/permutation significance testing between conditions of interest, applied identically to every compared condition. This history is surfaced not as a headline finding but as part of a full account of how the reported numbers were obtained, in a subfield where inconsistent or underspecified evaluation protocols are a plausible contributor to the widespread range of reported AUCs (Section 2); it is also, concretely, what allowed the structured-only finding above to be identified and reported with confidence rather than overlooked.

This study has several further limitations. First, and most consequential, it uses a single institution's cohort (UTSW-Glioma); no independent cohort was available for this study, and external validation (e.g., on TCGA, UCSF-PDGM, or UPenn-GBM) is needed before these AUCs – and, specifically, the finding that structured features alone match or exceed image-plus-structured fusion – can be treated as generalizable rather than as specific to this cohort's acquisition protocol and case mix. Second, the 3D CNN baseline is an in-house implementation of a standard fully-supervised, end-to-end architecture rather than a tuned, extensively optimized model in its own right, so the 0.723 baseline AUC should be read as a representative, not necessarily maximal, estimate of what an end-to-end 3D CNN can achieve on this cohort. Third, MGMT and 1p/19q codeletion results are based on substantially smaller labeled subsets (n=277 and n=140) than IDH and grade, and their wide confidence intervals should be read accordingly. Fourth, the explainability analysis presented here is correlational (attention weight vs. tumor fraction) and does not itself explain what the attention mechanism *is* responding to instead, nor does it fully explain why the pooled image representation fails to add value over structured features; these are related but

distinct open questions. Fifth, the tumor segmentation masks used throughout this study – for slice bagging and as the reference standard in the attention-tumor correspondence analysis – are produced by the FeTS automated segmentation pipeline distributed with the dataset (Section 3.1), not by manual expert delineation; their quality was not independently re-verified in this study (e.g., via Dice overlap against a manually corrected subset), so a fraction of the near-chance attention-tumor correlation reported in Section 4 could in principle reflect segmentation noise rather than the model failing to attend to the true tumor region. This is considered a secondary explanation at most, since FeTS-derived segmentations are widely used and generally reliable for whole-tumor delineation, but is flagged as an assumption the explainability conclusion depends on. Sixth, the label-efficiency and calibration analyses (Section 4) were conducted on the fusion model specifically and have not yet been repeated for the structured-only model that this study's own ablation identifies as the stronger IDH predictor; this asymmetry is a direct, near-term target for follow-up work (Section 5.5).

5.5 Future work

Immediate next steps are: (1) external validation on at least one additional public glioma MRI cohort, with particular attention to whether the structured-only-versus-fusion pattern reported here replicates; (2) repeating the label-efficiency and calibration analyses (Section 4) for the structured-only model, since it currently achieves the higher point estimate for IDH; (3) extending feature and slice extraction to include the peritumoral (edema) region alongside the tumor core, motivated directly by the substantial independent MGMT signal that peritumoral features carry in Yu et al. and Li et al. [10, 11], as the most concrete lead identified here for improving MGMT prediction specifically; (4) now that a structured-only control condition establishes a clear bar to beat, a controlled ablation of partial backbone fine-tuning and domain-specific self-supervised pretraining, to test directly whether either can produce an image branch that significantly exceeds structured features alone, rather than merely matching or trailing them as the frozen backbone does here; (5) validating attention weight against radiologist-marked regions of diagnostic interest, rather than only against the full tumor segmentation, to test whether attention tracks a clinically meaningful subregion even where it does not track the whole tumor; (6) adopting a stability-aware radiomic feature-selection pipeline – inter-reader ICC filtering followed by minimum-redundancy-maximum-relevance selection, in the style of Li et al. [19] – if multi-reader or manually revised segmentations become available for this cohort, to test whether a more conservative feature set changes the structured-only and fusion results; and (7) resampling all sequences to a common isotropic voxel spacing prior to radiomic feature extraction, to bring the pipeline into full IBSI compliance and remove the residual scanner/geometry dependence noted in Section 3.4.

6 Conclusion

This paper presents GlioMIL, a 2.5D MIL framework for non-invasive glioma molecular profiling built on a frozen, ImageNet-pretrained EfficientNetV2-S backbone, evaluated on the UTSW-Glioma cohort under a single, shared, patient-level 5-fold

cross-validation protocol with bootstrap confidence intervals and paired significance testing. A controlled five-way ablation shows that frozen-backbone transfer learning alone (IDH AUC 0.711–0.756) is broadly comparable to a fully-supervised 3D CNN baseline (0.723), and that fusing standardized clinical and radiomic features into the pooled MIL representation delivers a large, statistically significant gain over image-only MIL, raising IDH AUC to 0.847 and grade AUC to 0.751 ($p < 0.0001$ for both). Critically, however, a structured-only control condition using the same clinical and radiomic features with no image input at all reaches an IDH AUC of 0.881, significantly higher than the fused model ($p = 0.009$), and is statistically indistinguishable from fusion for grade ($p = 0.60$): the frozen image branch, as constructed here, is not demonstrated to add predictive value beyond structured features on this cohort. MGMT and 1p/19q codeletion remain near chance in every condition, most plausibly reflecting their much smaller labeled subsets. Beyond point estimates, the fusion model is shown to be reasonably well calibrated (ECE 0.060, with a large accuracy gap between confident and unconfident predictions) and largely label-efficient (saturating near 50% of available training labels), but its attention weights – often assumed in the MIL literature to serve as a spatial explanation – correspond to tumor location only at a near-chance level, a specific and quantitative caution for how such models' explanations are presented and used. The full evaluation protocol is reported, including an earlier flawed ensembling scheme and a feature-normalization leakage issue that were identified and corrected, in the interest of trustworthy and reproducible reporting for this task; that same rigor is what surfaced the structured-only finding as a central, rather than incidental, result of this study.

Acknowledgements. The author thanks the curators of the UTSW-Glioma dataset and TCIA for making this cohort publicly available for research use.

Abbreviations. The following abbreviations are used in this manuscript (Table 4).

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics approval and consent to participate. This study is a secondary, retrospective analysis of the UTSW-Glioma dataset, a fully de-identified, publicly available imaging and clinical dataset distributed through TCIA [20–22]. No new human subjects were recruited, and no identifiable personal data were accessed; the original data collection and de-identification were governed by the data-contributing institution's own ethics approval and consent process prior to public release. In accordance with TCIA's data usage policy and standard practice for secondary analysis of de-identified public research datasets, no additional institutional ethics approval was required for this study.

Consent for publication. Not applicable. This study uses only de-identified, publicly available imaging data and contains no identifiable individual patient information, images, or case details that would require patient consent for publication.

Availability of data and materials. The UTSW-Glioma dataset used in this study is publicly available via TCIA at https://doi.org/10.7937/dfae-1b86.

Table 4 List of abbreviations.

Abbreviation	Definition
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CI	Confidence Interval
CNN	Convolutional Neural Network
ECE	Expected Calibration Error
FeTS	Federated Tumor Segmentation
FLAIR	Fluid-Attenuated Inversion Recovery
GLCM	Gray-Level Co-occurrence Matrix
GLDM	Gray-Level Dependence Matrix
GLRLM	Gray-Level Run-Length Matrix
GLSZM	Gray-Level Size-Zone Matrix
IBSI	Image Biomarker Standardisation Initiative
IDH	Isocitrate Dehydrogenase
MC (dropout)	Monte Carlo (dropout)
MGMT	O^6-Methylguanine-DNA Methyltransferase
MIL	Multiple Instance Learning
MRI	Magnetic Resonance Imaging
OOF	Out-Of-Fold
STARD	Standards for Reporting of Diagnostic Accuracy Studies
T1CE	Post-Contrast T1-Weighted (MRI sequence)
TCIA	The Cancer Imaging Archive
TRIPOD+AI	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence
WHO	World Health Organization
1p/19q	Chromosome arms 1p and 19q (codeletion)

Code availability. The implementation (tumor-guided slice extraction, frozen-backbone feature extraction, attention-MIL, fusion, structured-only model training, cross-validation, bootstrap evaluation, paired significance testing, label-efficiency, calibration, and explainability analyses) was developed in PyTorch, torchvision, PyRadiomics, and SimpleITK, and run on Kaggle GPU kernels as described in Section 3. The code is available to editors and reviewers upon request and will be made publicly available upon acceptance.

Reporting guidelines. This manuscript's reporting of model development and evaluation follows, to the extent applicable to a retrospective, single-cohort methodological study, the TRIPOD+AI [28] and STARD [29] reporting guidelines for prediction-model and diagnostic-accuracy studies; a completed checklist is available from the corresponding author upon request and is intended for inclusion as supplementary material before publication.

Authors' contributions. T.A. conceived the study, curated and processed the dataset, designed and implemented the model architecture and evaluation protocol, ran all experiments, analysed the results, and wrote the manuscript. J.A., R.M., and O.A. critically reviewed the manuscript and contributed to its intellectual content. All authors read and approved the final manuscript.

Conflict of interest. The authors declare no conflict of interest.

References

[1] Price M, Ballard CAP, Benedetti JR, Neff C, Cioffi G, Waite KA, et al. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2018–2022. Neuro-Oncology. 2025;27(Supplement 4):iv1–iv66. https://doi.org/10.1093/neuonc/noaf194.

[2] Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology. 2021;23(8):1231–1251. https://doi.org/10.1093/neuonc/noab106.

[3] Kim M, Jung SY, Park JE, Jo Y, Park SY, Nam SJ, et al. Diffusion- and perfusion-weighted MRI radiomics model may predict isocitrate dehydrogenase (IDH) mutation and tumor aggressiveness in diffuse lower grade glioma. European Radiology. 2020;30:2142–2151. https://doi.org/10.1007/s00330-019-06548-3.

[4] Arita H, Kinoshita M, Kawaguchi A, Takahashi M, Narita Y, Terakawa Y, et al. Lesion location implemented magnetic resonance imaging radiomics for predicting IDH and TERT promoter mutations in grade II/III gliomas. Scientific Reports. 2018;8:11773. https://doi.org/10.1038/s41598-018-30273-4.

[5] Choi YS, Bae S, Chang JH, Kang SG, Kim SH, Kim J, et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. Neuro-Oncology. 2021;23:304–313. https://doi.org/10.1093/neuonc/noaa177.

[6] Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Scientific Reports. 2020;10:7733. https://doi.org/10.1038/s41598-020-64588-y.

[7] Niu W, Yan J, Hao M, Zhang Y, Li T, Liu C, et al. MRI transformer deep learning and radiomics for predicting IDH wild type TERT promoter mutant gliomas. npj Precision Oncology. 2025;9:89. https://doi.org/10.1038/s41698-025-00884-y.

[8] He J, Ren J, Niu G, Liu A, Wu Q, Xie S, et al. Multiparametric MR radiomics in brain glioma: models comparation to predict biomarker status. BMC Medical Imaging. 2022;22:137. https://doi.org/10.1186/s12880-022-00865-8.

[9] He A, Wang P, Zhu A, Liu Y, Chen J, Liu L. Predicting IDH Mutation Status in Low-Grade Gliomas Based on Optimal Radiomic Features Combined with Multi-Sequence Magnetic Resonance Imaging. Diagnostics. 2022;12(12):2995. https://doi.org/10.3390/diagnostics12122995.

[10] Yu X, Zhou J, Wu Y, Bai Y, Meng N, Wu Q, et al. Assessment of MGMT promoter methylation status in glioblastoma using deep learning features from

multi-sequence MRI of intratumoral and peritumoral regions. Cancer Imaging. 2024;24:172. https://doi.org/10.1186/s40644-024-00817-1.

[11] Li L, Xiao F, Wang S, Kuang S, Li Z, Zhong Y, et al. Preoperative prediction of MGMT promoter methylation in glioblastoma based on multiregional and multi-sequence MRI radiomics analysis. Scientific Reports. 2024;14:16031. https://doi.org/10.1038/s41598-024-66653-2.

[12] Ilse M, Tomczak JM, Welling M. Deep Multiple Instance Learning for Digital Histopathology. In: Handbook of Medical Image Computing and Computer Assisted Intervention. Elsevier; 2020. p. 521–546.

[13] Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. In: Proceedings of the 35th International Conference on Machine Learning. vol. 80. PMLR; 2018. p. 2127–2136.

[14] Jain S, Wallace BC. Attention is not Explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019. p. 3543–3556.

[15] Tan M, Le Q. EfficientNetV2: Smaller Models and Faster Training. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139. PMLR; 2021. p. 10096–10106.

[16] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 16000–16009.

[17] Coupet M, Urruty T, Leelanupab T, Naudin M, Bourdon P, Fernandez Maloigne C, et al. A multi-sequences MRI deep framework study applied to glioma classfication. Multimedia Tools and Applications. 2022;81(10):13563–13591. https://doi.org/10.1007/s11042-022-12316-1.

[18] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research. 2017;77(21):e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[19] Li J, Xu Q, Fan X, Cheng X, Tao J, Lu H, et al. Multi-Sequence MRI radiomics model for discrimination of recurrence and pseudoprogression in gliomas. European Journal of Radiology. 2026;194:112508. https://doi.org/10.1016/j.ejrad.2025.112508.

[20] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging. 2013;26(6):1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

[21] Reddy D, Saadat N, Holcomb J, Wagner B, Truong N, Bowerman J, et al.: The University of Texas Southwestern Glioma MRI dataset with molecular marker characterization and segmentations (UTSW-Glioma) (Version 1) [Data set]. The Cancer Imaging Archive. 625 patients treated at UTSW between 2006 and 2023.

[22] Reddy DD, Saadat N, Holcomb JM, Wagner BC, Truong NC, Bowerman J, et al. The University of Texas Southwestern Glioma Dataset – MRI, Molecular Markers and Segmentations. Scientific Data. 2026;13:934. https://doi.org/10.1038/s41597-026-07274-4.

[23] Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated Learning Enables Big Data for Rare Cancer Boundary Detection. Nature Communications. 2022;13:7346. https://doi.org/10.1038/s41467-022-33407-5.

[24] Zwanenburg A, Vallíeres M, Abdalah MA, Aerts HJWL, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338. https://doi.org/10.1148/radiol.2020191145.

[25] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845. https://doi.org/10.2307/2531595.

[26] Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Proceedings of the 33rd International Conference on Machine Learning. vol. 48. PMLR; 2016. p. 1050–1059.

[27] Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: Proceedings of the 34th International Conference on Machine Learning. vol. 70. PMLR; 2017. p. 1321–1330.

[28] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

[29] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. BMJ. 2015;351:h5527. https://doi.org/10.1136/bmj.h5527.