Journal of Imaging

MDPI

Article

Does Tumor-Mask Guidance Improve Glioma Grade Classification from T2-Weighted MRI Alone? An Out-of-Fold Evaluation with nnU-Net and Ground-Truth Localization

Vivek Sanker [1,*], Dhanya Mahesh [2], Zhikai Li [3], Alexander Thaller [4], Philip Heesen [5], Linda Liverani [1], David Wang [6], Maria Jose Cavagnaro [1], Ravi Teja Medikonda [1], Laura Prolo [1], Harminder Singh [1,7], John Ratliff [1], Atman Desai [1] and Michael Lim [1]

[1] Department of Neurosurgery, Stanford University, Stanford, CA 94304, USA; liverani@stanford.edu (L.L.); mjcava@stanford.edu (M.J.C.); rmediko1@stanford.edu (R.T.M.); lmprolo@stanford.edu (L.P.); harman@stanford.edu (H.S.); jratliff@stanford.edu (J.R.); atman@stanford.edu (A.D.); mklim@stanford.edu (M.L.)
[2] Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA; dhanya.mahesh.med@dartmouth.edu
[3] Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK; zl498@cam.ac.uk
[4] Department of Neurosurgery, Medical University of Graz, 8010 Graz, ST, Austria; alexander.thaller@medunigraz.at
[5] Department of Neurosurgery, University of Zurich, 8091 Zurich, Switzerland; philip.heesen@uzh.ch
[6] Faculty of Medicine, University of Oslo, 0316 Oslo, Norway; david.wang@studmed.uio.no
[7] Department of Neurosurgery, Santa Clara Valley Medical Center, 751 S Bascom Avenue, San Jose, CA 95128, USA
* Correspondence: vsanker@stanford.edu

Abstract

Multi-sequence MRI is the standard for glioma assessment, but complete or contrast-enhanced protocols are not always available (withheld gadolinium, low-field MRI, incomplete archives). We asked whether tumor localization improves grade classification from T2-weighted MRI alone. In this retrospective study of the BraTS2020 training cohort (369 patients; legacy labels 293 high-grade [HGG], 76 low-grade [LGG]), 74 patients (15 LGG) were held out at the patient level for internal testing. A 2D U-Net trained in five development folds supplied out-of-fold whole-tumor masks for training 3D CNN classifiers; nnU-Net served as a segmentation benchmark. Classifiers received T2 alone or T2 plus a mask, which at test time came from the 2D U-Net, nnU-Net, the ground truth (oracle) or random noise. Analyses used five seeds, DeLong tests with Holm correction and development-set cross-validation. Whole-tumor Dice was 0.807 (95% CI 0.773–0.840) for the 2D U-Net and 0.877 (0.857–0.895) for nnU-Net. The T2-only CNN reached an ROC-AUC of 0.814 (95% CI 0.702–0.913). Mask-guided variants reached 0.677–0.742, none above T2-only (Holm-adjusted $p = 1.00$ for all four confirmatory comparisons of informative masks). Cross-validation reproduced this ranking (0.782 vs. 0.707). With LGG as the positive class, PR-AUC was 0.571 (no-skill 0.203). Tumor-mask guidance did not improve T2-only grading at any level of mask quality, including ground truth. The results concern a legacy grade label in one public cohort and require external testing.

Keywords: glioma; T2-weighted MRI; grade classification; deep learning; U-Net; nnU-Net; out-of-fold prediction; missing modality; negative result

Check for updates

Academic Editor: Aly A. Farag

Received: 6 August 2026
Revised: 15 October 2026
Accepted: 17 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Gliomas are the most common malignant primary tumors of the brain. In the United States, the age-adjusted incidence of malignant brain and other central nervous system (CNS) tumors is 6.89 per 100,000 population per year; gliomas account for 22.9% of all primary CNS tumors, and glioblastoma alone for 51.5% of malignant ones [1]. The 2021 World Health Organization (WHO) classification assigns grades within molecularly defined tumor types; IDH-mutant astrocytomas, for example, span CNS WHO grades 2–4, and IDH-wildtype diffuse astrocytic gliomas with TERT promoter mutation, EGFR amplification or combined chromosome 7 gain and 10 loss are classified as glioblastoma, grade 4, even without high-grade histology [2]. The high-grade (HGG) versus low-grade (LGG) labels distributed with BraTS2020 predate this framework and derive from histopathological diagnoses of the contributing cohorts [3]; in this study they are treated as a legacy research label and not as a surrogate for an integrated WHO 2021 diagnosis. A preoperative estimate of tumor aggressiveness nevertheless bears on the urgency and extent of surgery and on adjuvant planning [4,5].

Multi-sequence MRI (native T1, contrast-enhanced T1 [T1ce], T2 and FLAIR) is the standard of care for preoperative glioma assessment, and T2 alone is not proposed here as a substitute for it. T2-weighted images nevertheless carry grade-related information of their own; the T2–FLAIR mismatch sign, for example, is a specific marker of IDH-mutant astrocytoma [6]. Most deep-learning pipelines built on the multimodal Brain Tumor Segmentation (BraTS) benchmark and on U-Net or nnU-Net architectures assume all four co-registered sequences [3,7–9]. Convolutional networks can also grade gliomas from conventional MRI [10], but reported performance is often inflated by evaluation against binarized rather than official multi-class labels, by ground-truth segmentations reaching the test set, and by the absence of strong baselines on identical patient splits [9,11,12]. Such leakage of information between training and test data is a recognized source of over-optimistic results in machine-learning-based science [13].

There are, however, specific situations in which only non-contrast sequences are available or reliable. Gadolinium-based contrast is used selectively: exposure during pregnancy has been associated with stillbirth or neonatal death and with inflammatory skin conditions in the child [14]; immediate allergic-like reactions occur in about 9 per 10,000 administrations [15]; gadolinium is retained in the brain after repeated doses, which argues for restraint in patients who need many follow-up examinations [16]; and in acute kidney injury or severe chronic kidney disease, the very low risk of nephrogenic systemic fibrosis has to be weighed against the harm of delaying or withholding contrast [17]. Portable low-field (0.064 T) scanners used at the bedside of critically ill patients acquire T1, T2, FLAIR and diffusion-weighted images without contrast and at lower signal-to-noise ratios [18], and in retrospective or multi-institutional archives, sequences are often missing or degraded by motion, which has motivated a dedicated line of missing-modality methods [19]. T2 is usually among the retained sequences. For such settings it is useful to know how much grade-related information a classifier can extract from T2 alone and whether it extracts more when it is told where the tumor is.

Explicit tumor localization is a plausible way to obtain more. Imaging features associated with high grade, such as necrosis and heterogeneous signal, lie within the lesion, which occupies a small part of the brain volume, and a classifier trained on a few hundred patients may not learn to find them unaided. With all four sequences, Pereira et al. graded BraTS 2017 gliomas with a 3D CNN more accurately from an automatically segmented tumor region than from the whole brain (accuracy 0.930 vs. 0.895) [20], and segmentation-derived regions of interest have been used to guide CNN classification of MGMT promoter methylation [21]. Without T1ce, however, the mask must itself be predicted from T2, on which

the enhancing tumor, defined on T1ce [12], is not directly visible. Whether a T2-derived mask still adds information to a T2-only classifier, and whether its effect depends on mask quality, has to our knowledge not been examined. Two-stage designs of this kind also carry a specific risk of optimistic evaluation: if the masks of training patients are predicted by a segmentation network that was trained on the same patients, the classifier learns from masks of better quality than those it receives at test time.

We therefore addressed three research questions (RQ), using the T2 channel of BraTS2020 as a controlled test bed in which labels, preprocessing and patient split are held fixed (Figure 1). RQ1: How accurately can the whole tumor and its sub-regions be segmented from T2 alone by a 2D U-Net and by nnU-Net [9]? RQ2: Does a whole-tumor mask, added as a second input channel, improve discrimination of the legacy HGG/LGG label by a T2-only 3D CNN? RQ3: Does any effect depend on the quality of the mask, from the 2D U-Net to nnU-Net to the reference segmentation? On the basis of the multi-sequence results above, we expected the mask to improve discrimination and the improvement to increase with mask quality. To prevent the train–test mismatch described above, classifiers were trained with out-of-fold 2D U-Net masks; at test time, the mask channel was supplied by the 2D U-Net, by nnU-Net or by the reference segmentation, with a random-noise channel as negative control. Each condition was trained with five seeds and compared with the T2-only model by paired tests with Holm correction, and the comparison was repeated by cross-validation in the development set.

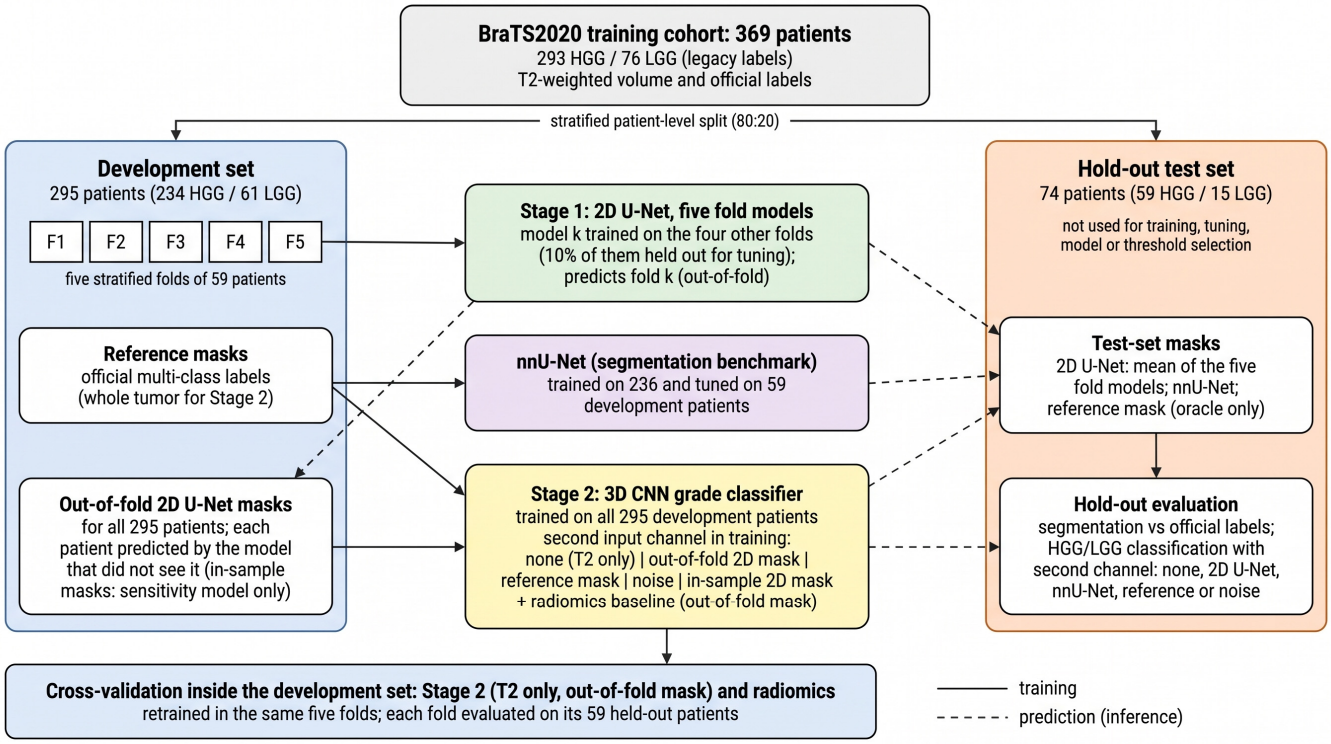


Figure 1. Experimental design. The 369 BraTS2020 patients were split once at the patient level into a development set (blue) and a hold-out test set (orange). Solid arrows connect each model with the data it was trained on; dashed arrows show the data to which it was applied. The five Stage-1 2D U-Net fold models supply out-of-fold masks for Stage-2 training and a five-model ensemble mask for the test set; nnU-Net, trained on the development set, supplies an alternative test-set mask. Stage-2 3D CNNs were trained on all development patients with T2 alone or T2 plus a second channel and evaluated once on the test set; cross-validation repeats Stage-2 training within the development folds. Training settings of each model are given in Table 1.

Table 1. Data used for training, tuning and testing of each model, with the main training settings (patient numbers in parentheses). Folds are the five stratified folds of the 295-patient development set; the test set (74 patients) was used only for the final evaluation. Channel 2, second input channel of the 3D CNN; OOF, out-of-fold 2D U-Net mask; GT, ground-truth mask; CE, cross-entropy; lr, learning rate. Complete settings are given in Supplementary Table S1 and Supplementary Methods S2 and S3.

Model	Training Data	Tuning	Testing	Settings
Stage 1: 2D U-Net, 5-fold models	4 folds minus 10% (212 patients)	24 patients of the training folds: lr schedule, checkpoint	Held-out fold (59; OOF masks); test set (74; ensemble)	Dice + CE; AdamW, lr 1×10^{-3}; batch 16; 30 epochs
nnU-Net 3d_fullres	236 development patients	59 development patients; monitoring only	Test set (74)	nnU-Net defaults; 250 epochs
Stage 2: 3D CNN, hold-out (5 conditions $\times$ 5 seeds)	295 development patients; channel 2: none, OOF, GT, noise or in-sample mask	None; final epoch	Test set (74); channel 2: 2D ensemble, nnU-Net, GT or noise	Focal loss; AdamW, lr 1×10^{-4}; batch 4; 30 epochs
Stage 2: 3D CNN, cross-validation (2 conditions $\times$ 3 seeds)	4 folds (236 patients); channel 2: none or OOF mask	None; final epoch	Held-out fold (59)	As hold-out
Radiomics logistic regression	295 development patients or 4 folds; OOF mask	None	Test set (74) or held-out fold	L2 (C = 1); balanced class weights

2. Materials and Methods

2.1. Study Design and Dataset

This retrospective study used the training cohort of the BraTS2020 challenge (369 patients: 293 HGG, 76 LGG) [3,7,12]; its reporting follows the CLAIM 2024 checklist [22] (Supplementary Materials). BraTS2020 consists of routine preoperative MRI examinations acquired with different clinical protocols and scanners at multiple institutions and distributed in de-identified form after co-registration to a common template, resampling and skull stripping [12]. Of the 369 patients, 167 originate from The Cancer Genome Atlas glioma collections of The Cancer Imaging Archive (TCGA/TCIA) [3], 129 from the Center for Biomedical Image Computing and Analytics (CBICA, University of Pennsylvania), 30 from the BraTS 2013 data and 43 from other contributing institutions. All 369 patients were included; none was excluded. The reference standard for segmentation was the official BraTS label map (necrotic/non-enhancing tumor, peritumoral edema, enhancing tumor), segmented manually by one to four raters following a common protocol, approved by experienced neuroradiologists and reviewed for consistency by a board-certified neuroradiologist [12]; it is referred to below as the ground truth (GT). Individual rater annotations and inter-rater agreement are not distributed. The reference standard for classification was the HGG/LGG label distributed with BraTS2020, which derives from the histopathological diagnoses of the contributing institutions (Section 1); HGG was coded as the positive class.

Patients were split once into a development set (n = 295; 234 HGG, 61 LGG) and a hold-out test set for internal testing (n = 74; 59 HGG, 15 LGG) by a stratified patient-level 80:20 split (scikit-learn train_test_split, random_state = 42). Labels of the official BraTS2020 validation and test cohorts are not public, so no established labeled split was available; the 80:20 ratio is conventional and left 15 LGG patients for testing, and the test-set size was not determined by a power calculation. The two partitions were similar in source (TCGA/TCIA 134/295 [45%] vs. 33/74 [45%]) and, among the 236 HGG patients for whom age is distributed, in age (median 61.0 vs. 63.4 years; Mann–Whitney $p = 0.28$); sex is not distributed with BraTS2020. The test set was not used for training, tuning, model selection,

early stopping or threshold choice. The development set was divided into five stratified folds (StratifiedKFold, shuffled, random_state = 42), which were used both for Stage-1 out-of-fold prediction and for cross-validation of Stage 2, so that the Stage-1 mask of a validation patient always came from a model that had not seen that patient.

2.2. Image Preprocessing

Only the T2-weighted volume was used. BraTS2020 volumes are co-registered, resampled to 1 mm isotropic resolution and skull-stripped by the challenge organizers ($240 \times 240 \times 155$ voxels); the T2 series were originally acquired as axial 2D sequences with 2–4 mm slice thickness [12], and per-patient acquisition parameters are not distributed. Each volume was min–max scaled to [0, 1]; no bias-field correction, inter-patient intensity standardization or outlier removal was applied, and no image data were missing. Stage 1 used all 155 axial slices (240×240) of each patient. For Stage 2, the T2 volume and the localization mask were resampled to $128 \times 128 \times 96$ voxels by trilinear interpolation. The Stage-1 2D U-Net was trained without augmentation; Stage-2 training used random flips along both in-plane axes ($p = 0.5$ each) and multiplicative intensity scaling of the T2 channel by a factor drawn from U(0.9, 1.1). No test-time augmentation was used for either model. The identity of the T2 channel was verified for every converted file type (Supplementary Methods S4; Supplementary Figure S3).

2.3. Stage 1: 2D U-Net Tumor Localization

Stage 1 was a single-channel 2D U-Net [8] with four 2×2 max-pooling levels (blocks of two 3×3 convolutions with batch normalization and ReLU followed by spatial dropout of 0.1; transposed-convolution up-sampling with skip connections; base width 32; 7.76 M parameters; Supplementary Figure S1) predicting the four official BraTS classes (background, necrotic/non-enhancing tumor, edema, enhancing tumor; label 4 remapped to 3) on axial slices. It was trained with 0.5·Dice + 0.5·cross-entropy loss, AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4}), batch size 16, mixed precision and 30 epochs, with ReduceLROnPlateau (factor 0.5, patience 3) on the loss of an inner validation (tuning) subset. Weights were initialized with the PyTorch 2.7.1 (CUDA 11.8) defaults; no pretrained weights were used. In each development fold, the model was fitted on the other four folds minus a random 10% inner validation subset (212 fitting and 24 inner-validation patients), and the checkpoint with the highest inner-validation mean Dice over whole tumor (WT; labels 1–3), tumor core (TC; labels 1 and 3) and enhancing tumor (ET; label 3) was kept (training curves in Supplementary Figure S4). This model predicted the 59 patients of the held-out fold, yielding out-of-fold (OOF) label maps for all 295 development patients, analogous to the out-of-fold predictions used in stacked generalization [23]. Test-set masks were obtained by averaging the softmax outputs of the five fold models. To quantify how much better masks are for patients that a model has seen, each development patient was also predicted in-sample by a fold model whose fitting set contained that patient; these masks were used only for the sensitivity model (v) in Section 2.4. Official BraTS2020 segmentations were the reference for all patients except BraTS20_Training_355, for which no official file was available in the downloaded archive, and the label map was rebuilt from the HDF5 archive; rebuilt and official maps were voxel-identical in all 368 patients with both (Supplementary Methods S1).

2.4. Stage 2: Mask-Guided 3D Grade Classification

The Stage-2 classifier was a 3D CNN (four blocks of $3 \times 3 \times 3$ convolution, batch normalization and ReLU with 32–256 filters, the first three followed by $2 \times 2 \times 2$ max pooling; adaptive average pooling to $4 \times 4 \times 4$; fully connected layers of 512 and 128 units with ReLU and dropout 0.5; 9.6 M parameters; PyTorch default initialization without

pretraining; Supplementary Figure S2) with a 128 × 128 × 96 input of one channel (T2-only) or two channels (T2 plus a binary WT mask). Conditions differed only in the second channel: (i) T2-only; (ii) mask-guided, trained with Stage-1 OOF masks and tested with the five-model ensemble masks; (iii) mask-guided, trained with ground-truth masks; (iv) a random-mask control whose second channel was patient-fixed uniform noise in [0, 1]; and (v) a sensitivity model trained with in-sample Stage-1 masks and tested with the ensemble masks, which exposes the classifier to better masks in training than at test. To separate the effect of mask quality from the effect of having a mask, models (ii) and (iii) were also evaluated with the test-time channel replaced by nnU-Net masks (50- and 250-epoch checkpoints; Section 2.5), by the 2D ensemble masks or by the ground-truth mask; ground truth at test time is an oracle reference and is not available in deployment. All models used class-weighted focal loss ($\gamma = 2$; class weight $n/(2 \cdot n_c)$), AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-4}), cosine annealing with warm restarts ($T_0 = 10$, T_mult = 2), batch size 4, mixed precision, shuffled (not class-balanced) sampling and 30 epochs. Training length was fixed at 30 epochs without early stopping, and the final-epoch model was used; no checkpoint was selected on validation or test data. Each condition was trained with five seeds (42–46), and the primary prediction is the seed-averaged probability of HGG, thresholded at 0.5. Table 1 summarizes the training, tuning and test data and the training settings of every model; all implemented settings are listed in Supplementary Table S1 and Supplementary Methods S3.

2.5. nnU-Net Segmentation Benchmark

nnU-Net v2 [9] (version 2.6.2; 3d_fullres configuration with default planning and preprocessing; T2 as the only input channel; official multi-class labels) was trained on the development set only, using nnU-Net's default fold-0 split of the 295 development patients (236 training, 59 validation). Because of the compute budget, training ran for 250 instead of the default 1000 epochs, and a snapshot at epoch 50 was kept as an under-trained reference (Supplementary Methods S2). Both checkpoints, neither selected on validation performance, predicted the 74 test patients with default test-time mirroring. nnU-Net served as a segmentation benchmark for Stage 1 and as a higher-quality test-time localization channel for Stage 2 (Section 2.4).

2.6. Secondary Analyses

Cross-validation. Stage 2 was also evaluated by five-fold cross-validation within the development set, reusing the Stage-1 folds so that the masks of each validation fold were out-of-fold. T2-only and mask-guided (OOF-mask) models were trained with three seeds (42–44) and the recipe of Section 2.4, and a logistic regression on five first-order features of the T2 signal inside the 2D mask (volume, mean, standard deviation, maximum, 90th percentile; standardized; balanced class weights) was fitted in the same folds. Interpretability. Integrated Gradients [24] (zero baseline, 32-step right Riemann sum, target = HGG logit; implemented in PyTorch) was computed for all 74 test patients with the seed-42 T2-only and mask-guided models. Attribution enrichment was defined as the fraction of absolute T2-channel attribution inside the ground-truth WT divided by the WT fraction of brain voxels, so that 1 indicates no spatial preference for the tumor. The attributions were not validated against an independent reference and are interpreted descriptively. Timing. Inference time per volume was measured on an otherwise idle NVIDIA RTX 5000 Ada GPU for 19 test patients after one warm-up case, separately for reading and scaling, 2D U-Net inference (single model and five-model ensemble) and resampling plus 3D CNN inference; nnU-Net was timed from the difference between 10-case and 1-case runs to exclude model loading.

2.7. Statistical Analysis

Segmentation was scored per patient against the official multi-class labels with Dice and the 95th-percentile Hausdorff distance (HD95, mm) for WT, TC and ET. Following the BraTS convention, Dice = 1 and HD95 = 0 when both prediction and reference were empty, and HD95 = 373.1 mm (image diagonal) when only one was empty. Classification metrics were accuracy, balanced accuracy, sensitivity (HGG), specificity (LGG), weighted F1, ROC-AUC and PR-AUC (average precision), with the Brier score as a summary of calibration. PR-AUC is reported for both choices of positive class because its no-skill value equals the prevalence of the positive class, 0.797 with HGG and 0.203 with LGG as positive (score $1 - P(HGG)$) [25]. Confidence intervals (95%) were obtained by patient-level bootstrap with 2000 resamples, stratified by grade for classification metrics; segmentation models were compared on the same test patients by the bootstrap of the paired mean difference and by the Wilcoxon signed-rank test. Each condition was compared with T2-only on the same test patients by the DeLong test for correlated ROC curves [26,27], by paired bootstrap of the ROC-AUC and balanced-accuracy differences, and by the exact McNemar test at the 0.5 threshold. The confirmatory family consisted of five comparisons with T2-only (OOF-trained mask-guided model with 2D and with nnU-Net-250 test masks; GT-trained model with ground-truth and with nnU-Net-250 test masks; random-mask control), and Holm correction [28] was applied within this family; all other comparisons are descriptive. Cross-validation results are summarized as the mean over folds with a t-based 95% CI and as metrics of the pooled out-of-fold predictions, on which paired tests were performed. Analyses used Python 3.13.9 with PyTorch 2.7.1 (CUDA 11.8), nnU-Net 2.6.2, NumPy 2.2.6, SciPy 1.15.2, scikit-learn 1.7.2 and nibabel 5.3.3.

2.8. Pipeline Verification

Each stage of the pipeline consumes the output of the previous one, so an error in one stage propagates to the next without being visible in the final metrics. The following properties were therefore verified before the final analysis. (i) Sequence identity: the BraTS2020 HDF5 archive stores the sequences in the order FLAIR, T1, T1ce, T2, and in the NIfTI files converted from it, the file named *_T2 contains T1ce; the T2 volume was identified from image contrast in every file type and used throughout (Supplementary Methods S4; Supplementary Figure S3). (ii) Reference labels: Stage 1 and nnU-Net were trained and evaluated on the official multi-class labels on axial slices, and the label map rebuilt from the HDF5 archive was compared voxel-wise with the official files (Supplementary Methods S1). (iii) Patient separation: the patient lists of all training, tuning and test sets were compared, and no test patient entered the training of any model, including the segmentation networks that supply the masks. (iv) Model selection: no epoch, checkpoint, threshold or hyperparameter was chosen on test data. (v) Class imbalance: imbalance was corrected once through the loss weights; adding a class-balanced sampler to the weighted loss shifted predictions toward LGG without improving ROC-AUC (Supplementary Table S5). (vi) Cross-validation was confined to the development set. A preliminary version of this pipeline, whose results were archived as version 1 of the Zenodo record, did not meet these criteria: it used the T1ce file as T2, trained Stage 1 on binary necrotic-core masks on non-axial slices, included test patients in segmentation training, selected the classifier epoch on the test set, corrected class imbalance twice and cross-validated over all 369 patients. Its results are superseded and are not used in this article; the corrections are listed in the version history of the archive.

2.9. Generative AI Use

ChatGPT 5.6 was used for code development/debugging. The authors take full responsibility for the content of this paper.

3. Results

3.1. Segmentation Performance on Multi-Class Ground Truth

On the 74 test patients, the five-model 2D U-Net ensemble reached mean Dice of 0.807 (0.773–0.840) for WT, 0.595 (0.535–0.649) for TC and 0.378 (0.324–0.429) for ET, with a WT HD95 of 11.5 (8.6–14.8) mm (Table 2; Figures 2 and 3). nnU-Net was better in every region: 0.849/0.707/0.450 after 50 epochs and 0.877/0.717/0.466 after 250 epochs (WT HD95 6.0 (4.5–8.1) mm). The paired Dice differences between nnU-Net-250 and the 2D ensemble were 0.069 (0.048–0.094) for WT, 0.122 (0.088–0.159) for TC and 0.088 (0.053–0.127) for ET (Wilcoxon $p < 0.001$ for each). Out-of-fold Dice in the development set (0.797/0.586/0.358; per-fold values in Supplementary Table S4) was close to the test-set values, so the masks used to train Stage 2 resemble those it receives at test time. In-sample predictions for the same development patients were markedly better (WT 0.864 vs. 0.797, TC 0.782 vs. 0.586, ET 0.497 vs. 0.358; Wilcoxon $p < 0.001$ for each region), which quantifies the train–test mismatch that in-sample masks introduce. ET was the weakest region for all models, as expected without the T1ce sequence on which enhancement is defined.

Table 2. Hold-out segmentation from T2 alone (official multi-class BraTS labels). Mean Dice and mean HD95 (95th-percentile Hausdorff distance) with patient-level bootstrap 95% CIs (2000 resamples). ET was absent in 6 test patients; ET HD95 means are dominated by the 373 mm penalty applied when exactly one of prediction and reference is empty (medians 10.1, 5.8 and 5.6 mm for the three test-set models). The last row gives the out-of-fold development-set masks used to train Stage 2. WT, whole tumor; TC, tumor core; ET, enhancing tumor.

Method	Dice WT	Dice TC	Dice ET	HD95 WT (mm)	HD95 TC (mm)	HD95 ET (mm)	n
2D U-Net, five-fold ensemble (test)	0.807 (0.773–0.840)	0.595 (0.535–0.649)	0.378 (0.324–0.429)	11.5 (8.6–14.8)	12.6 (10.5–14.9)	40.5 (20.4–65.4)	74
nnU-Net 3d_fullres, 50 epochs (test)	0.849 (0.824–0.872)	0.707 (0.656–0.758)	0.450 (0.388–0.507)	9.2 (6.5–12.2)	13.8 (7.6–25.0)	43.0 (18.9–68.3)	74
nnU-Net 3d_fullres, 250 epochs (test)	0.877 (0.857–0.895)	0.717 (0.665–0.769)	0.466 (0.404–0.527)	6.0 (4.5–8.1)	8.3 (6.8–10.0)	43.6 (22.2–68.9)	74
2D U-Net, out-of-fold (development)	0.797 (0.778–0.814)	0.586 (0.560–0.614)	0.358 (0.333–0.383)	15.7 (13.8–17.8)	24.6 (19.1–31.2)	49.1 (37.8–61.5)	295

The gap between the 2D ensemble and nnU-Net-250 was largest for TC (0.12) and ET (0.09) and smallest for WT (0.07). Stage 2 uses only the WT mask, for which the average difference is small but concentrated in a subset of patients with poor 2D localization (Figure 3).

Median WT Dice was 0.877 for the 2D ensemble and 0.908 for nnU-Net-250, but the 2D model had a longer lower tail: 9 test patients had WT Dice below 0.6, compared with 4 for nnU-Net-50 and 1 for nnU-Net-250. These are the patients in whom a better localization channel would be expected to matter most for Stage 2.

Figure 4a illustrates the overlap of the legacy classes in T2 appearance: the LGG example (ID 274) contains labeled necrotic and enhancing components, whereas the HGG example (ID 252) has a compact core with surrounding edema.

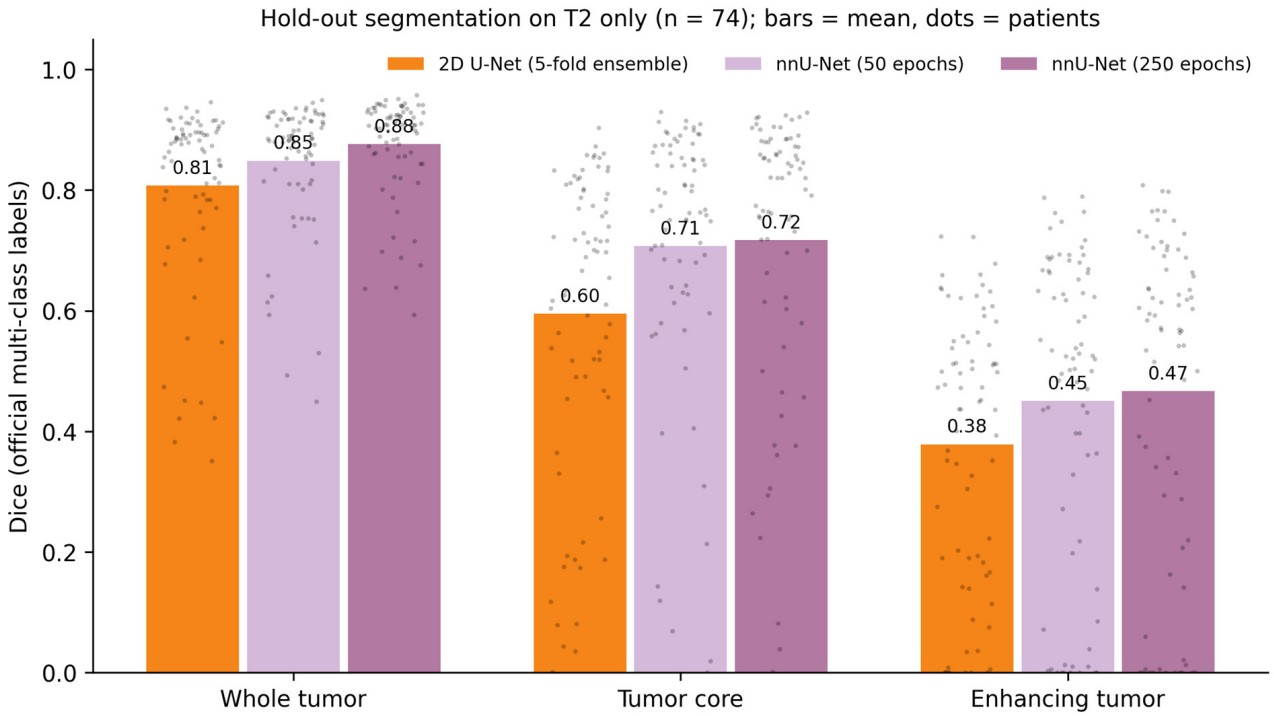


Figure 2. Regional Dice on the hold-out set (n = 74) for the 2D U-Net ensemble and nnU-Net after 50 and 250 epochs. Bars show means; dots show individual patients.

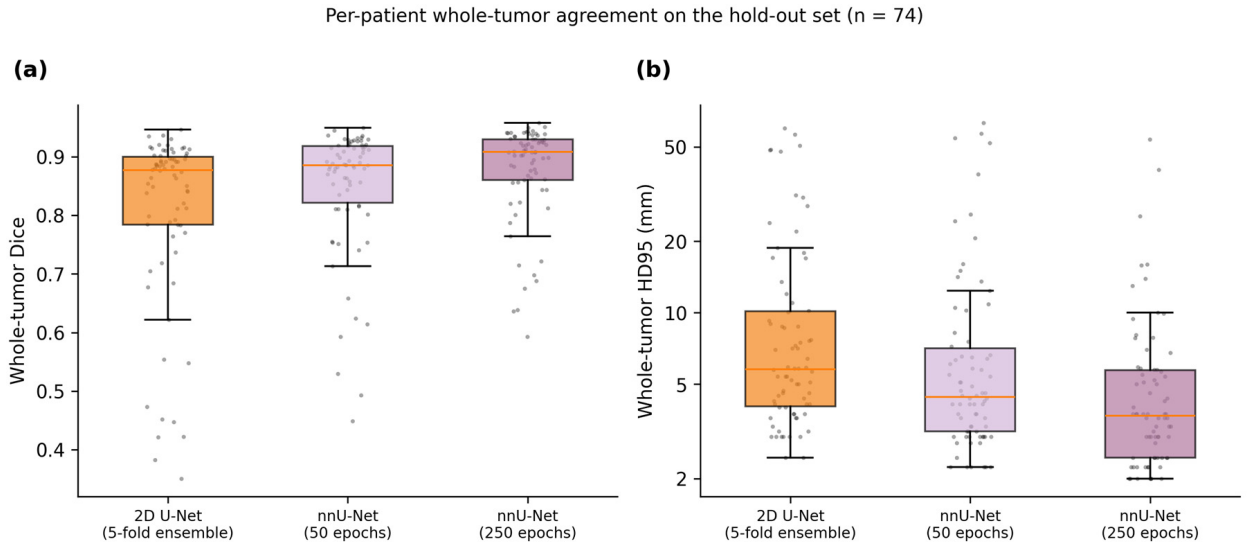


Figure 3. Per-patient whole-tumor Dice (**a**) and HD95 ((**b**); logarithmic axis) on the hold-out set for the 2D U-Net ensemble and nnU-Net after 50 and 250 epochs. Boxes show the interquartile range with the median as a horizontal line and Tukey whiskers; gray dots are individual patients.

ID 274 is the worst-segmented test patient for the 2D ensemble (WT Dice 0.35), whereas nnU-Net recovers most of the lesion (0.88); in ID 252, both models are close to the reference (0.88 and 0.94) (Figure 4b).

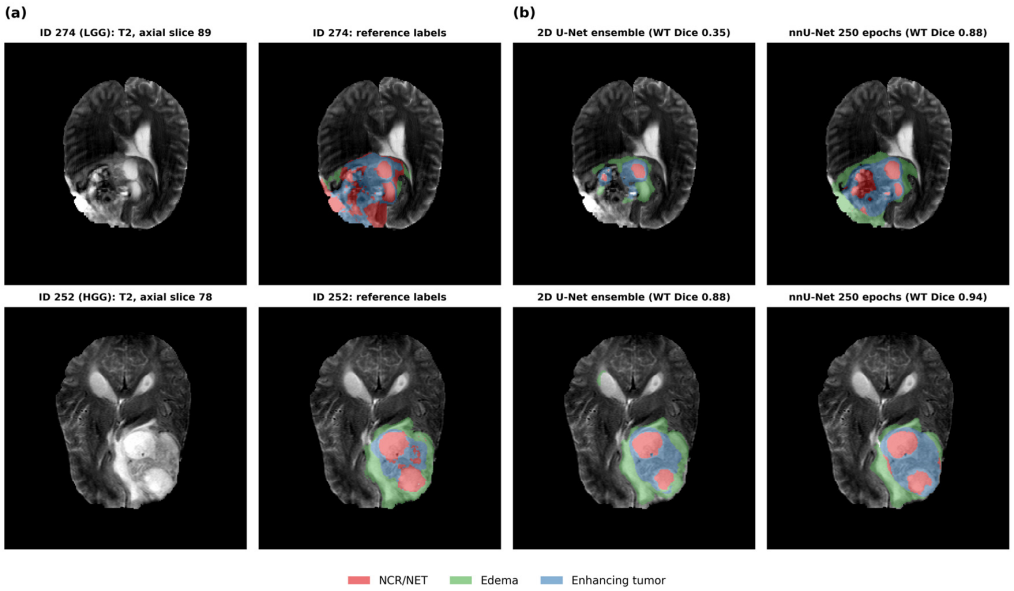


Figure 4. Two hold-out patients, ID 274 (legacy LGG; axial slice 89) and ID 252 (legacy HGG; slice 78), each shown at the slice with the largest tumor area. (**a**) T2-weighted input and official multi-class labels. (**b**) Predicted segmentations of the 2D U-Net five-model ensemble and of nnU-Net after 250 epochs, with the per-patient WT Dice in each panel title. Red, necrotic/non-enhancing tumor; green, edema; blue, enhancing tumor. Display window: 1st–99th intensity percentiles.

3.2. Classification Performance

Table 3 and Figures 5–7 summarize hold-out classification. The T2-only 3D CNN had the highest ROC-AUC of all conditions (0.814, 95% CI 0.702–0.913) and the highest balanced accuracy (0.747). Adding the 2D U-Net mask lowered ROC-AUC to 0.714 (ΔAUC −0.099, DeLong $p = 0.339$, Holm-adjusted $p = 1.00$) and balanced accuracy to 0.707. Accuracy was higher with the mask (0.770 vs. 0.676) mainly because the mask-guided model predicted HGG more often (sensitivity 0.814 vs. 0.627; specificity 0.600 vs. 0.867). The ΔAUC was negative for each of the five seeds (range −0.177 to −0.068). The random-mask control fell to chance (ROC-AUC 0.474; Holm-adjusted $p < 0.001$) and, after seed averaging, assigned every patient to LGG. Training with in-sample instead of OOF masks gave ROC-AUC 0.728 and balanced accuracy 0.624, so neither mask source offered an advantage over T2-only. The first-order radiomics baseline was worse than the T2-only CNN (ROC-AUC 0.664; DeLong $p = 0.041$, descriptive).

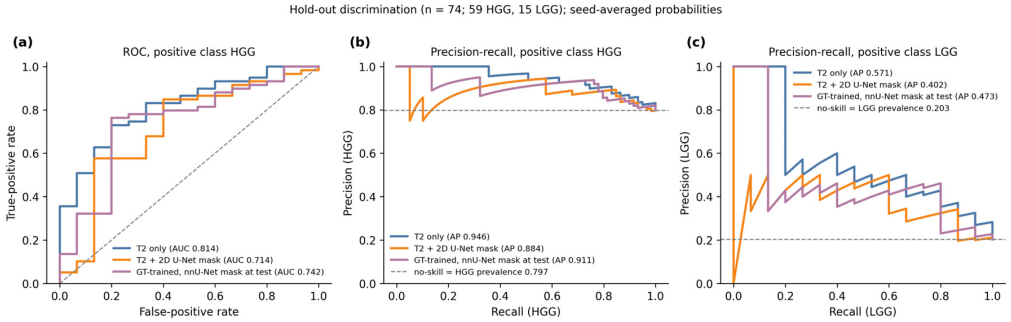


Figure 5. Hold-out discrimination for the T2-only model, the mask-guided model with 2D U-Net masks and the GT-trained model tested with nnU-Net-250 masks (seed-averaged probabilities). (**a**) ROC curves with HGG as positive class; dashed line, chance (AUC 0.5). (**b**) Precision–recall with HGG as positive class; dashed line, no-skill level 0.797. (**c**) Precision–recall with LGG as positive class; dashed line, no-skill level 0.203.

Table 3. Hold-out classification of the legacy BraTS2020 HGG/LGG label (n = 74; 59 HGG, 15 LGG). Values are for seed-averaged predictions (five seeds; radiomics is a single deterministic model), with stratified bootstrap 95% CIs for ROC-AUC. PR-AUC is given with HGG as positive class (no-skill 0.797) and with LGG as positive class (no-skill 0.203). Last column: ΔAUC versus T2-only with DeLong p; for the five confirmatory comparisons, the Holm-adjusted p is given after the semicolon. Mask sources are listed as training/test; GT, ground truth (oracle when used at test); OOF, out-of-fold.

Model (Mask: Training/Test)	Accuracy	Balanced Acc.	ROC-AUC (95% CI)	PR-AUC HGG+/LGG+	ΔAUC (p; Holm p)
T2-only 3D CNN	0.676	0.747	0.814 (0.702–0.913)	0.946/0.571	reference
Mask-guided: 2D OOF/2D ensemble	0.770	0.707	0.714 (0.547–0.866)	0.884/0.402	−0.099 (p = 0.339; 1.00)
Mask-guided: 2D OOF/nnU-Net-250	0.743	0.715	0.713 (0.542–0.871)	0.880/0.417	−0.101 (p = 0.371; 1.00)
Mask-guided: 2D OOF/GT	0.784	0.715	0.686 (0.498–0.852)	0.860/0.379	−0.128 (p = 0.280)
Mask-guided: GT/GT (oracle)	0.743	0.615	0.714 (0.550–0.861)	0.893/0.385	−0.099 (p = 0.333; 1.00)
Mask-guided: GT/nnU-Net-250	0.757	0.723	0.742 (0.588–0.879)	0.911/0.473	−0.071 (p = 0.469; 1.00)
Mask-guided: 2D in-sample/2D ensemble	0.757	0.624	0.728 (0.581–0.863)	0.906/0.450	−0.086 (p = 0.273)
Random-mask control	0.203	0.500	0.474 (0.307–0.639)	0.780/0.212	−0.340 (p < 0.001; <0.001)
Radiomics logistic regression (2D mask)	0.473	0.620	0.664 (0.515–0.807)	0.864/0.295	−0.149 (p = 0.041)

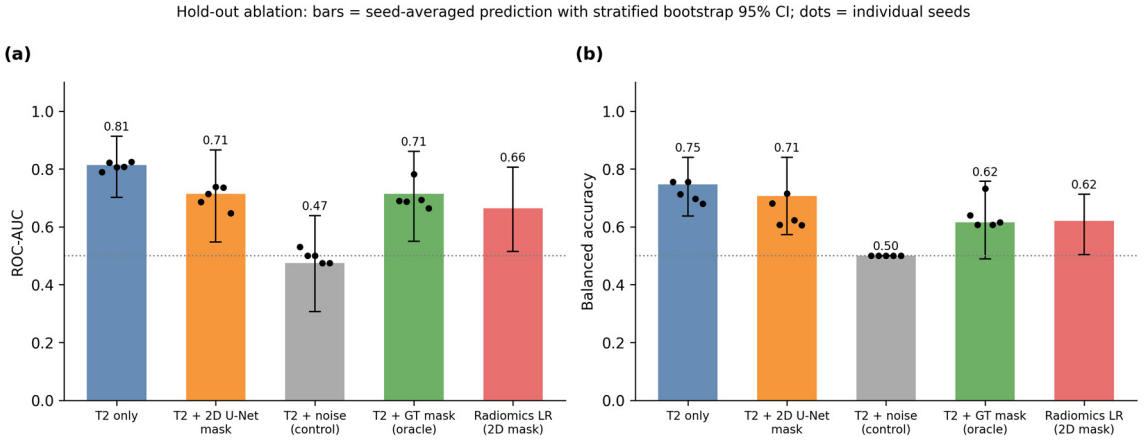


Figure 6. Hold-out ablation: ROC-AUC (**a**) and balanced accuracy (**b**) of the seed-averaged prediction with stratified bootstrap 95% CI; dots show the five individual seeds; dotted line, chance level.

With HGG as the positive class, the no-skill level of PR-AUC is the HGG prevalence (59/74 = 0.797), so the values of 0.946 (T2-only) and 0.884 (mask-guided) are moderate gains over chance. With LGG as positive (no-skill 0.203), PR-AUC was 0.571 for T2-only and 0.402 for the mask-guided model (Figure 5b,c). The LGG-positive curve is the more informative one here, because recognition of the minority class is where the models differ.

At the 0.5 threshold, the T2-only model classified 37 of 59 HGG and 13 of 15 LGG test patients correctly, the mask-guided model 48 and 9; the two models agreed on 53 patients (43 correct, 10 incorrect). Most errors of the T2-only model were HGG patients predicted as LGG (22 of 24). These patients had larger tumors than correctly classified HGG patients (median ground-truth WT volume, including edema, 158 vs. 66 mL; Mann–Whitney

$p = 0.001$), in the range of the test LGG patients (median 117 mL), whereas their 2D U-Net WT Dice did not differ (median 0.89 vs. 0.87; $p = 0.495$). The same volume pattern held for the mask-guided model (170 vs. 65 mL; $p < 0.001$). Misclassified HGG patients came more often from the TCGA/TCIA collections (11/22 vs. 9/37 correctly classified; Fisher $p = 0.053$). Calibration was similar for the two models (Brier score 0.199 for T2-only and 0.217 for the mask-guided model).

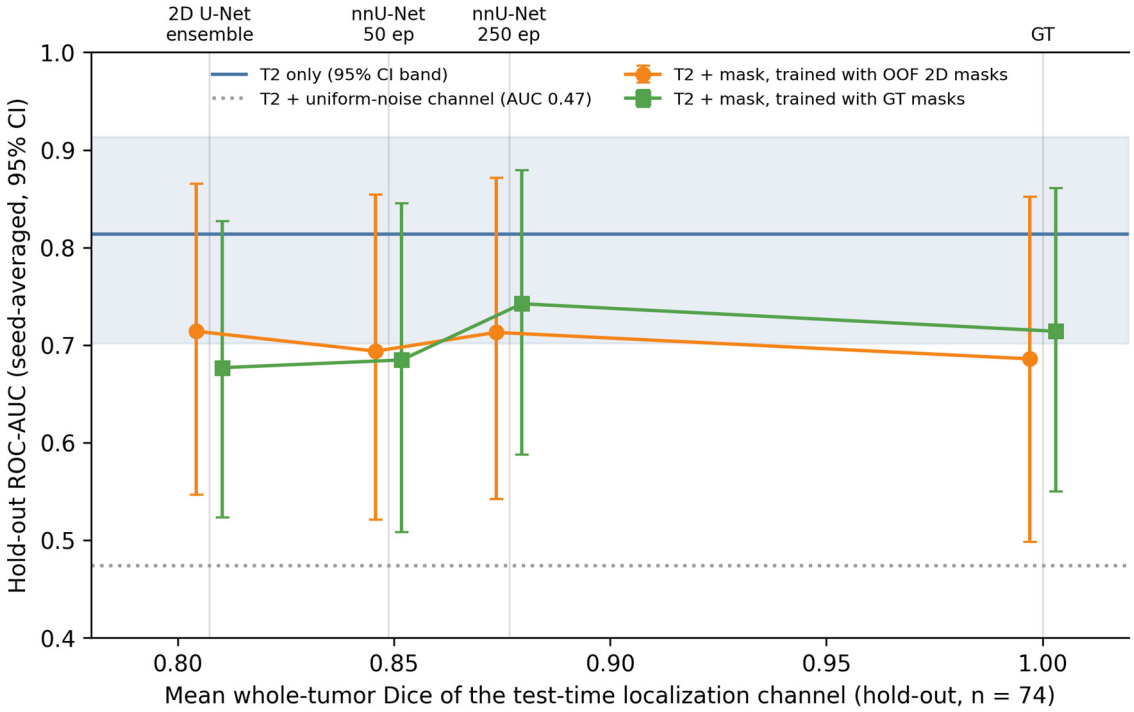


Figure 7. Hold-out ROC-AUC as a function of the quality of the test-time localization channel (mean WT Dice of the 2D ensemble, nnU-Net-50, nnU-Net-250 and ground truth) for models trained with OOF 2D masks (orange) and with ground-truth masks (green). Error bars, bootstrap 95% CI; blue line and band, T2-only estimate and 95% CI; dotted line, random-mask control.

The seed-level results in Figure 6 show that the ranking does not depend on one favorable seed: per-seed ROC-AUC ranged from 0.789 to 0.825 for T2-only, from 0.647 to 0.738 for the mask-guided model and from 0.664 to 0.782 for the GT-trained oracle (Supplementary Table S3). The random-mask control shows that an uninformative second channel is harmful; the informative channels, however, did not add discrimination over T2 alone.

Figure 7 addresses whether grading depends on localization quality. Raising mask quality from WT Dice 0.81 (2D ensemble) to 0.88 (nnU-Net-250) and 1.0 (ground truth) did not raise ROC-AUC for the model trained on OOF masks (0.714, 0.713 and 0.686). The GT-trained model reached its highest value with nnU-Net-250 masks (0.742) and 0.714 with ground truth, still below T2-only (ΔAUC -0.071 and -0.099; Holm-adjusted $p = 1.00$ for both). Within the range of localization quality tested, improving the segmentation mask did not improve grading performance. Even when ground-truth masks were provided, the mask-guided models did not exceed the T2-only model in point-estimate ROC-AUC. The mask channel was not ignored: it received about half of the attribution in the mask-guided model (Section 3.4). Taken together, these findings suggest that, within the tested architecture and mask-concatenation strategy, neither improved localization quality nor the addition of an informative spatial prior provided measurable benefit over T2 alone. However, the confidence intervals remain compatible with smaller effects and do not exclude potential benefit from localization information under other architectures or fusion

strategies. In contrast, an uninformative second channel (random-mask control, ROC-AUC 0.474) was harmful.

3.3. Cross-Validation Robustness

Five-fold cross-validation within the development set (n = 295; Figure 8; Supplementary Table S2) reproduced the hold-out ranking. Fold-mean ROC-AUC was 0.782 (95% CI 0.701–0.862) for T2-only, 0.707 (0.607–0.806) for the mask-guided model with OOF masks and 0.641 (0.528–0.754) for radiomics; balanced accuracy was 0.704, 0.672 and 0.600. The mask-guided model was below T2-only in four of five folds, and on the pooled out-of-fold predictions its ROC-AUC was 0.715 versus 0.770 (ΔAUC −0.054, 95% CI −0.132 to 0.023; DeLong $p = 0.169$). Radiomics was worse than T2-only (ΔAUC −0.139; $p = 0.002$).

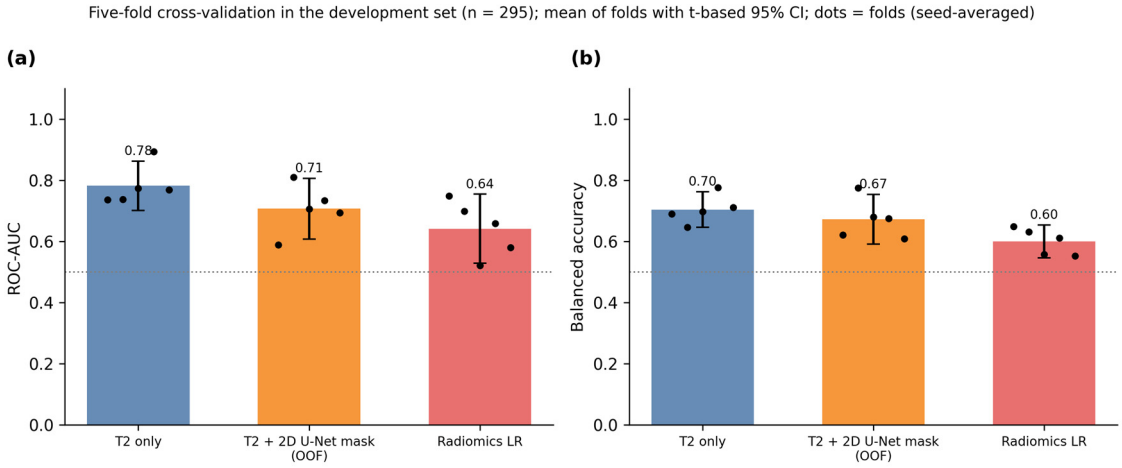


Figure 8. Five-fold cross-validation within the development set (n = 295): (**a**) ROC-AUC and (**b**) balanced accuracy. Bars, mean over folds with t-based 95% CI; dots, individual folds (CNN predictions averaged over three seeds).

Hold-out evaluation and cross-validation therefore agree in direction. ROC-AUC values were slightly lower in cross-validation, where each Stage-2 model was trained on 236 rather than 295 patients.

3.4. Interpretability

Integrated Gradients did not show tumor-focused attribution (Figure 9). The share of absolute T2-channel attribution inside the ground-truth WT was smaller than the WT share of brain volume in most patients: median enrichment 0.63 (IQR 0.54–0.82) for T2-only and 0.48 (0.37–0.69) for the mask-guided model, with enrichment above 1 in 15% and 7% of patients. In the mask-guided model, about half of the total attribution (median 51%) fell on the mask channel itself. Both classifiers therefore draw on T2 appearance across the brain rather than on tumor voxels specifically.

3.5. Computational Performance

On an idle RTX 5000 Ada GPU (19 test volumes), reading and scaling a T2 volume took 0.16 ± 0.01 s, one 2D U-Net 0.45 ± 0.03 s (five-model ensemble 1.28 ± 0.03 s) and resampling plus 3D CNN classification 0.15 ± 0.01 s; the T2-only classifier thus needs about 0.3 s per volume including reading. nnU-Net-250 inference with default test-time mirroring took about 2.9 s per case. These are model-inference times on preprocessed BraTS volumes. They exclude DICOM conversion, co-registration, skull stripping and reporting, which dominate end-to-end processing in clinical practice, and they do not demonstrate clinical feasibility.

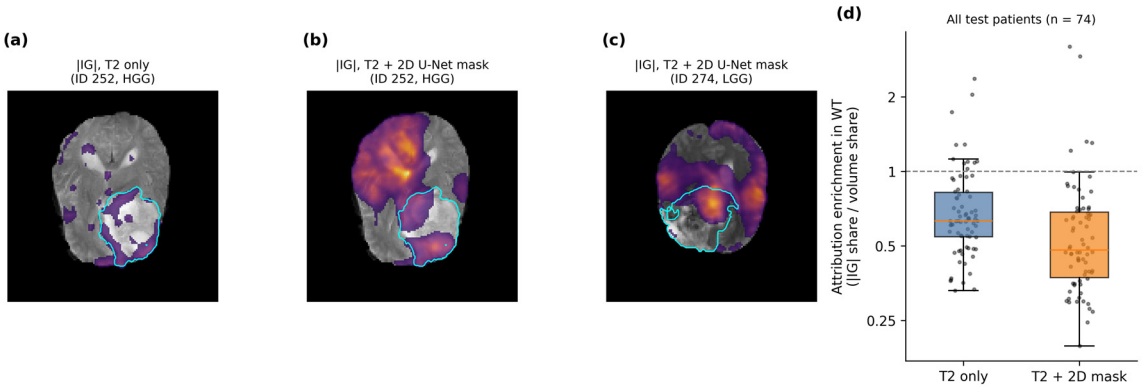


Figure 9. Integrated Gradients: absolute attribution to the T2 channel (smoothed for display; cyan, ground-truth WT contour) for (**a**) the T2-only and (**b**) the mask-guided model in ID 252 (HGG) and (**c**) the mask-guided model in ID 274 (LGG); (**d**) attribution enrichment in WT for all 74 test patients (1 = no spatial preference; logarithmic axis). (**a**–**c**) grayscale, T2 slice; purple-to-yellow overlay, absolute attribution (higher = yellow); cyan, ground-truth WT contour. (**d**) boxes, interquartile range with median (horizontal line) and Tukey whiskers; gray dots, individual patients; dashed line, enrichment = 1 (no spatial preference for the tumor); logarithmic axis.

4. Discussion

The main result is negative. On the T2 channel of BraTS2020, a 3D CNN given only the T2 volume discriminated the legacy HGG/LGG label better, in point estimate, than every mask-guided variant tested, including one trained and tested with ground-truth masks, and cross-validation in the development set gave the same ranking. No difference between an informative mask condition and T2-only was significant after Holm correction, so the data do not show that localization guidance is harmful; with five seeds and two evaluation protocols, they do show that it does not help in this setting. This differs from the multi-sequence study of Pereira et al., in which grading from an automatically segmented tumor region was more accurate than grading from the whole brain [20]; that design cropped the input to the tumor region instead of adding a mask channel, and it included T1ce, on which enhancement and the necrotic core are delineated. In two-stage designs, the source of the training masks also matters: in-sample masks of the development patients were markedly better than out-of-fold masks (WT Dice 0.864 vs. 0.797), so a classifier trained on them sees better masks in training than at test. Here, this mismatch did not change the conclusion (ROC-AUC 0.728 with in-sample masks), but out-of-fold masks are needed for training and test inputs to come from the same distribution.

Several explanations for the absent benefit are compatible with the data, and none was tested directly. The Integrated Gradients analysis indicates that the T2-only network does not rely mainly on tumor voxels, so a channel that marks the tumor adds little that the network does not already use, while it doubles the input size for a training set of 295 patients (61 LGG). The binary WT mask also encodes tumor extent and shape, which overlap considerably between the legacy classes (compare IDs 274 and 252 in Figure 4a), and it discards the graded T2 signal inside the lesion. Designs that use the mask to crop or weight the T2 input, rather than appending it as a channel, may behave differently and remain to be tested.

The clinical settings in which T2-only analysis is relevant are restricted (Section 1): withheld gadolinium, bedside low-field imaging, and incomplete or degraded archival examinations. The present data address these settings only indirectly, because BraTS2020 consists of preprocessed high-field examinations from which the T2 volume was extracted; performance on native low-field or motion-degraded T2 is likely to be lower and has to be measured on such data. What the study provides is an estimate of the grade-related signal

available to a T2-only model under controlled conditions (ROC-AUC 0.81 on hold-out and 0.77–0.78 in cross-validation) and evidence that tumor masks, even ground-truth masks, do not raise it.

From T2 alone, WT Dice was 0.81 for the 2D ensemble and 0.88 for nnU-Net, close to the WT Dice of multi-sequence U-Net models but clearly below them for TC and ET (Table 4), as expected when the sequence that defines enhancement is missing. The T2 Dice of 0.79 reported by Montaha et al. [29] is not directly comparable, because it refers to a binary tumor mask on the middle axial slice of each volume rather than to a volumetric WT mask. For whole-tumor localization, a 2D U-Net therefore gives a usable approximation in about 1.3 s per volume (ensemble), whereas nnU-Net is preferable whenever sub-regional delineation matters.

Table 4 places these results next to representative U-Net studies on BraTS data; differences in tumor regions, input sequences and evaluation protocols preclude a direct ranking. A recent review of U-Net-based brain tumor segmentation identifies computational efficiency, generalization across sites and scanners, interpretability and clinical validation as open requirements for clinical use [30]. The present study examines only some of these, on public research data, and does not establish readiness for routine clinical use, which would require validation on non-BraTS MRI acquired and preprocessed in routine practice.

Table 4. Comparison with representative U-Net-based glioma segmentation studies on BraTS data. Tumor regions, input sequences and evaluation protocols differ, so the table provides context rather than a ranking. Values for this study are hold-out results on the 74-patient test set. ED, peritumoral edema.

Study	Method	Dataset	Reported Dice	Notes
Baid et al. 2020 [31]	3D U-Net	BraTS 2018 (validation set)	WT 0.88, TC 0.83, ET 0.75	T1, T1ce, T2 and FLAIR
Montaha et al. 2023 [29]	2D U-Net, one model per sequence	BraTS 2020	Binary tumor mask: 0.94 (T1), 0.91 (FLAIR), 0.79 (T2)	Single sequences; middle axial slice only, 128 × 128
Tene-Hurtado et al. 2022 [32]	2D U-Net/Attention U-Net	BraTS 2020	ED 0.85, ET 0.70, TC 0.74/ED 0.88, ET 0.72, TC 0.81	T1, T1ce, T2 and FLAIR
This study	2D U-Net ensemble/nnU-Net	BraTS 2020	WT 0.81, TC 0.60, ET 0.38/WT 0.88, TC 0.72, ET 0.47	T2 only; grading ROC-AUC 0.814 (T2-only 3D CNN), not improved by mask guidance

4.1. Limitations

Several limitations apply. First, the outcome is the legacy BraTS2020 HGG/LGG label derived from pre-2021 histological diagnoses, which does not map onto WHO 2021 integrated diagnoses [2] (Section 1). Neither model can therefore be read as predicting a contemporary diagnosis, and this limits clinical translation more than any performance figure reported here. Second, all data come from one public cohort without external testing, and BraTS volumes are centrally co-registered, skull-stripped and resampled, so performance on routine or low-field clinical T2 is unknown [33]. Sex is not distributed with BraTS2020 and age only for HGG patients, so performance could not be assessed in demographic subgroups. Third, the test set contains only 15 LGG patients; bootstrap CIs for ROC-AUC span about 0.2 to 0.35, and only large differences between conditions could have been detected. The upper 95% confidence limits of ΔAUC for the four informative confirmatory comparisons were 0.10–0.12, so a benefit of mask guidance larger than about 0.1 in ROC-AUC is unlikely, but smaller benefits or harms cannot be excluded. Fourth, nnU-Net was trained for 250 rather than 1000 epochs on one fold, so its Dice values are

a conservative estimate of nnU-Net performance on T2. Fifth, only one fusion strategy (concatenation of a binary WT mask) and one classifier architecture were tested. Finally, reported times cover model inference only.

4.2. Future Directions

Three directions follow. The T2-only grade signal should be measured on external clinical data, ideally in the settings that motivate T2-only analysis (non-contrast and low-field examinations), with calibration and uncertainty estimates [33]. Localization may be more useful when it restricts or weights the image input, for example through tumor-centered crops or attention mechanisms, than when appended as a channel; the out-of-fold protocol used here can be applied to such designs [20,21]. Where sub-regional delineation is needed, nnU-Net-class segmentation should be used [9]. Intraoperative applications, such as integration with 5-aminolevulinic acid fluorescence guidance, are not supported by the present data [34]. Finally, under the WHO 2021 framework, molecular targets such as IDH status are more relevant prediction targets than the legacy grade label [35], and future T2-based models should be developed on cohorts with integrated diagnoses.

5. Conclusions

On the T2 channel of BraTS2020, a single-channel 3D CNN discriminated the legacy HGG/LGG label with a hold-out ROC-AUC of 0.81 (balanced accuracy 0.75). Adding a tumor-localization channel did not improve discrimination, whether the mask came from an out-of-fold 2D U-Net, from nnU-Net or from the ground truth, and cross-validation in the development set gave the same ranking. T2-only segmentation reached WT Dice 0.81 with a 2D U-Net ensemble and 0.88 with nnU-Net. With a single public cohort, 15 LGG test patients and a legacy outcome label, these results provide a baseline for T2-only grading rather than a clinically usable tool.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jimaging12110533/s1: Supplementary Methods S1: Reference labels; Supplementary Methods S2: nnU-Net training; Supplementary Methods S3: Classification training and conditions; Supplementary Methods S4: Verification of the T2 channel; Table S1: Computational parameters; Table S2: Per-fold results of five-fold cross-validation within the development set (n = 295; folds identical to the Stage-1 folds). CNN predictions are averaged over three seeds (42–44); radiomics is deterministic. Pooled OOF = metric of all 295 out-of-fold predictions; Table S3: Hold-out ROC-AUC (n = 74) of each training seed and of the seed-averaged prediction for all Stage-2 conditions. Full metric sets with bootstrap CIs are in results/classification_stats.json of the Zenodo archive (DOI 10.5281/zenodo.23050626); Table S4: Stage-1 2D U-Net segmentation by development fold (out-of-fold predictions of the 59 held-out patients per fold), pooled OOF, in-sample predictions for the same 295 patients, and the test-set ensemble. In-sample versus OOF: Wilcoxon signed-rank $p < 0.001$ for WT, TC and ET; Table S5: Sensitivity of the T2-only hold-out model to class-imbalance handling (seeds 42–45; true T2; final epoch). All reported models correct imbalance once, through the class-weighted loss; adding a class-balanced sampler shifts predictions toward LGG without improving ROC-AUC. The test set contains 15 LGG patients; Figure S1: 2D U-Net encoder–decoder architecture for slice-wise segmentation. The network has one T2 input channel and four output classes; Figure S2: 3D CNN architecture for HGG/LGG classification (one input channel for T2-only, two for mask-guided variants). Schematic only: volumes are resampled to 128 × 128 × 96 before classification; labels such as ~240 × 240 × 155 in the drawing indicate native volume size; Figure S3: Verification of the T2 channel in two patients (IDs 1 and 274): the same axial slice from each converted file, labeled with the sequence identified from image contrast, and the official label. The file *_FLAIR.nii.gz contains the T2 volume used in this study; *_T2.nii.gz contains the T1ce volume; Figure S4: Stage-1 2D U-Net training in the five development folds: inner-validation mean Dice (WT, TC, ET; left) and inner-validation loss (right). A transient drop near epoch 10 at the initial learning

rate recovered after the plateau scheduler halved the learning rate; the retained checkpoints are those with the highest inner-validation Dice (epochs 27, 22, 8, 8, 30 for folds 1–5); Figure S5: nnU-Net (3d_fullres, T2 only) training on development fold 0 over 250 epochs, as logged by nnU-Net: training and validation loss with pseudo-Dice (top), epoch duration (middle) and learning rate (bottom); CLAIM 2024 checklist.

Author Contributions: Conceptualization, V.S.; methodology, V.S., D.M. and Z.L.; validation, V.S.; data curation, V.S.; writing—original draft preparation, V.S., D.M., Z.L., A.T., P.H., L.L. and D.W.; writing—review and editing, R.T.M., M.J.C., L.P., H.S., J.R. and A.T.; visualization, L.L.; supervision, A.D. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study as all data used are de-identified and publicly available.

Informed Consent Statement: Informed consent was waived as all data used are publicly available and de-identified.

Data Availability Statement: BraTS2020 data are available through the MICCAI BraTS challenge (https://www.med.upenn.edu/cbica/brats2020/, accessed on 17 October 2025). Analysis code is publicly available at https://github.com/vsanker21/-2D-U-Net-based-Brain-Tumor-Segmentation-and-Grade-Classification-Segmentation (accessed on 17 October 2025). The code, per-patient segmentation metrics, per-patient classifier probabilities for every condition and seed, and statistical results are archived at Zenodo as version 2.0.0 (DOI: 10.5281/zenodo.23050626; all versions: DOI 10.5281/zenodo.22086674). Version 1 of that record (DOI: 10.5281/zenodo.22086675) contains the results of the preliminary pipeline version described in Section 2.8 and is superseded. Model weights are available from the corresponding author on reasonable request.

Acknowledgments: During the preparation of this study, the authors used ChatGPT 5.6 for the purposes of code development and debugging. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Price, M.; Ballard, C.; Benedetti, J.; Neff, C.; Cioffi, G.; Waite, K.A.; Kruchko, C.; Barnholtz-Sloan, J.S.; Ostrom, Q.T. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2017–2021. *Neuro Oncol.* **2024**, *26*, vi1–vi85. [CrossRef]
2. Louis, D.N.; Perry, A.; Wesseling, P.; Brat, D.J.; Cree, I.A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.K.; Pfister, S.M.; Reifenberger, G.; et al. The 2021 WHO Classification of Tumors of the Central Nervous System: A summary. *Neuro Oncol.* **2021**, *23*, 1231–1251. [CrossRef]
3. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117. [CrossRef]
4. Molinaro, A.M.; Hervey-Jumper, S.; Morshed, R.A.; Young, J.; Han, S.J.; Chunduru, P.; Zhang, Y.; Phillips, J.J.; Shai, A.; Lafontaine, M.; et al. Association of Maximal Extent of Resection of Contrast-Enhanced and Non–Contrast-Enhanced Tumor With Survival Within Molecular Subgroups of Patients With Newly Diagnosed Glioblastoma. *JAMA Oncol.* **2020**, *6*, 495–503. [CrossRef]
5. Stupp, R.; Mason, W.P.; van den Bent, M.J.; Weller, M.; Fisher, B.; Taphoorn, M.J.B.; Belanger, K.; Brandes, A.A.; Marosi, C.; Bogdahn, U.; et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* **2005**, *352*, 987–996. [CrossRef]
6. Patel, S.H.; Poisson, L.M.; Brat, D.J.; Zhou, Y.; Cooper, L.; Snuderl, M.; Thomas, C.; Franceschi, A.M.; Griffith, B.; Flanders, A.E.; et al. T2–FLAIR Mismatch, an Imaging Biomarker for IDH and 1p/19q Status in Lower-grade Gliomas: A TCGA/TCIA Project. *Clin. Cancer Res.* **2017**, *23*, 6078–6085. [CrossRef]
7. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef]

8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
9. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
10. Yang, Y.; Yan, L.-F.; Zhang, X.; Han, Y.; Nan, H.-Y.; Hu, Y.-C.; Hu, B.; Yan, S.-L.; Zhang, J.; Cheng, D.-L.; et al. Glioma Grading on Conventional MR Images: A Deep Learning Study With Transfer Learning. *Front. Neurosci.* **2018**, *12*, 804. [CrossRef]
11. Varoquaux, G.; Cheplygina, V. Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digit. Med.* **2022**, *5*, 48. [CrossRef]
12. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv* **2018**, arXiv:1811.02629. [CrossRef]
13. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef]
14. Ray, J.G.; Vermeulen, M.J.; Bharatha, A.; Montanera, W.J.; Park, A.L. Association Between MRI Exposure During Pregnancy and Fetal and Childhood Outcomes. *JAMA* **2016**, *316*, 952–961. [CrossRef]
15. Behzadi, A.H.; Zhao, Y.; Farooq, Z.; Prince, M.R. Immediate Allergic Reactions to Gadolinium-based Contrast Agents: A Systematic Review and Meta-Analysis. *Radiology* **2018**, *286*, 471–482. [CrossRef]
16. Gulani, V.; Calamante, F.; Shellock, F.G.; Kanal, E.; Reeder, S.B. Gadolinium deposition in the brain: Summary of evidence and recommendations. *Lancet Neurol.* **2017**, *16*, 564–570. [CrossRef]
17. Weinreb, J.C.; Rodby, R.A.; Yee, J.; Wang, C.L.; Fine, D.; McDonald, R.J.; Perazella, M.A.; Dillman, J.R.; Davenport, M.S. Use of Intravenous Gadolinium-based Contrast Media in Patients with Kidney Disease: Consensus Statements from the American College of Radiology and the National Kidney Foundation. *Radiology* **2021**, *298*, 28–35. [CrossRef]
18. Sheth, K.N.; Mazurek, M.H.; Yuen, M.M.; Cahn, B.A.; Shah, J.T.; Ward, A.; Kim, J.A.; Gilmore, E.J.; Falcone, G.J.; Petersen, N.; et al. Assessment of Brain Injury Using Portable, Low-Field Magnetic Resonance Imaging at the Bedside of Critically Ill Patients. *JAMA Neurol.* **2021**, *78*, 41–47. [CrossRef]
19. Havaei, M.; Guizard, N.; Chapados, N.; Bengio, Y. HeMIS: Hetero-Modal Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; Volume 9901, pp. 469–477. [CrossRef]
20. Pereira, S.; Meier, R.; Alves, V.; Reyes, M.; Silva, C.A. Automatic Brain Tumor Grading from MRI Data Using Convolutional Neural Networks and Quality Assessment. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11038, pp. 106–114. [CrossRef]
21. Koska, İ.Ö.; Koska, Ç. Deep learning classification of MGMT status of glioblastomas using multiparametric MRI with a novel domain knowledge augmented mask fusion approach. *Sci. Rep.* **2025**, *15*, 3273. [CrossRef]
22. Tejani, A.S.; Klontzas, M.E.; Gatti, A.A.; Mongan, J.T.; Moy, L.; Park, S.H.; Kahn, C.E.; CLAIM 2024 Update Panel; Abbara, S.; Afat, S.; et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol. Artif. Intell.* **2024**, *6*, e240300. [CrossRef]
23. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
24. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328. Available online: https://proceedings.mlr.press/v70/sundararajan17a.html (accessed on 5 August 2026).
25. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
26. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef]
27. Sun, X.; Xu, W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Process. Lett.* **2014**, *21*, 1389–1393. [CrossRef]
28. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
29. Montaha, S.; Azam, S.; Rakibul Haque Rafid, A.K.M.; Hasan, M.Z.; Karim, A. Brain Tumor Segmentation from 3D MRI Scans Using U-Net. *SN Comput. Sci.* **2023**, *4*, 386. [CrossRef]
30. Bastug Koc, A.; Akgun, D. A Review of U-Net Based Deep Learning Frameworks for MRI-Based Brain Tumor Segmentation. *Diagnostics* **2026**, *16*, 506. [CrossRef]

31. Baid, U.; Talbar, S.; Rane, S.; Gupta, S.; Thakur, M.H.; Moiyadi, A.; Sable, N.; Akolkar, M.; Mahajan, A. A Novel Approach for Fully Automatic Intra-Tumor Segmentation with 3D U-Net Architecture for Gliomas. *Front. Comput. Neurosci.* **2020**, *14*, 10. [CrossRef]
32. Tene-Hurtado, D.; Almeida-Galárraga, D.A.; Villalba-Meneses, G.; Alvarado-Cando, O.; Cadena-Morejón, C.; Salazar, V.H.; Orozco-López, O.; Tirado-Espín, A. Brain Tumor Segmentation Based on 2D U-Net Using MRI Multi-modalities Brain Images. In *Smart Technologies, Systems and Applications*; Communications in Computer and Information Science; Springer International Publishing: Cham, Switzerland, 2022; Volume 1532, pp. 345–359. [CrossRef]
33. Finlayson, S.G.; Subbaswamy, A.; Singh, K.; Bowers, J.; Kupke, A.; Zittrain, J.; Kohane, I.S.; Saria, S. The Clinician and Dataset Shift in Artificial Intelligence. *N. Engl. J. Med.* **2021**, *385*, 283–286. [CrossRef]
34. Stummer, W.; Pichlmeier, U.; Meinel, T.; Wiestler, O.D.; Zanella, F.; Reulen, H.-J.; ALA-Glioma Study Group. Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: A randomised controlled multicentre phase III trial. *Lancet Oncol.* **2006**, *7*, 392–401. [CrossRef]
35. Chang, K.; Bai, H.X.; Zhou, H.; Su, C.; Bi, W.L.; Agbodza, E.; Kavouridis, V.K.; Senders, J.T.; Boaro, A.; Beers, A.; et al. Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging. *Clin. Cancer Res.* **2018**, *24*, 1073–1081. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.