

Structured Modality Dropout for Robust MRI-Based IDH Mutation Classification in Glioma

Mohisha Mongia

Vivekananda Institute of Professional Studies - Technical Campus

Piyush Gupta

University School of Automation and Robotics

Satya P Singh

`satya.prakash@nsut.ac.in`

Netaji Subhas University of Technology

Research Article

Keywords:

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11032826/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Structured Modality Dropout for Robust MRI-Based IDH Mutation Classification in Glioma

Mohisha Mongia¹, Piyush Gupta², and Satya P Singh^{3,*}

¹ *Vivekananda Institute of Professional Studies – Technical Campus, New Delhi, India,
12717702724_cse@vipstc.edu.in*

² *University School of Automation and Robotics, New Delhi, India,
piyush.6019051624@std.ggsipu.ac.in*

³ *Netaji Subhas University of Technology, New Delhi, India, satya.prakash@nsut.ac.in*

**Corresponding author: satya.prakash@nsut.ac.in*

Abstract

The classification of IDH gene mutation is an essential marker in the treatment of a group of brain tumors called glioma, indicating the favorability of prognosis and response to treatment. Current methods determine the IDH mutation status invasively through postoperative tissue and DNA analysis. It is favorable to predict the mutation status non-invasively from MRI scans to determine suitable surgical strategies. Existing MRI-based prediction models often assume the availability of all four MRI modalities. However, in real-world clinical settings, one or more modalities may be missing due to reasons such as image quality and inconsistent protocols. This severely impacts the performance of MRI-based deep learning models. This study proposes a structured modality dropout training strategy to improve the robustness of a 3D ResNet18 model for IDH prediction. We compared three training strategies: Baseline (no dropout), Dropout-30 (moderate), and Dropout-50 (aggressive). All three models were evaluated using five-fold stratified cross-validation across six realistic modality availability scenarios (all four, missing T1CE, missing FLAIR, missing T2, T1 and T2 only, T1 only). Performance was evaluated using area under the receiver operating characteristic curve (AUC-ROC), balanced accuracy, sensitivity, specificity, Brier score, and bootstrap confidence intervals (CI).

Key Findings: The Baseline model performed best when all four modalities were available (AUC 0.823 ± 0.042), but the Dropout models showed more stable performance as the number of missing modalities increased. The Dropout-30 model showed the best calibration, with the lowest Brier score of 0.198. Overall, Dropout-30 was the most balanced and clinically viable model, remaining the most stable across scenarios, best calibrated, and showing good generalization across scanner models.

1 Introduction

Gliomas are a heterogeneous group of brain tumors. They are one of the most common types of malignant primary central nervous system (CNS) tumors. The discovery of mu-

tations in isocitrate dehydrogenase (IDH) has significantly changed the identification and classification of gliomas. Gliomas with mutant IDH have a better response to therapy and a more positive prognosis compared to gliomas with wildtype IDH [1]. The fifth edition of the World Health Organization (WHO) Classification of Tumors of the Central Nervous System describes the absence or presence of the IDH gene as one of the essential findings in the diagnosis of glioma [2, 3]. IDH status is normally determined postoperatively through invasive techniques such as tumor DNA and tissue analysis [4]. However, it is desirable to know the IDH mutation status before resection in order to determine suitable surgical strategies. For this purpose, techniques such as magnetic resonance spectroscopy, radiomics, and rapid IDH gene mutation prediction have been developed [5–7]. However, these techniques face challenges such as high diagnostic cost, lack of standardization across laboratories, and limited feasibility. This motivates the development of non-invasive imaging-based techniques to predict molecular characteristics [8].

MRI includes four modalities: T1-Weighted (T1), T2-Weighted (T2), T1-Weighted with Contrast Enhancement (T1CE), and Fluid Attenuation Inversion Recovery (FLAIR). Deep learning models using these four modalities have demonstrated promising results in non-invasive IDH mutation classification [9–11]. Many methods developed using all four common modalities have shown significant success, but performance often drops when one or more modalities are unavailable. In real-world clinical scenarios, modalities may be unavailable for many reasons, such as low image quality, inconsistent protocols across labs, allergic reactions, or hardware failure [12, 13]. Fig. 1 shows representative axial slices across the four modalities used in this study.

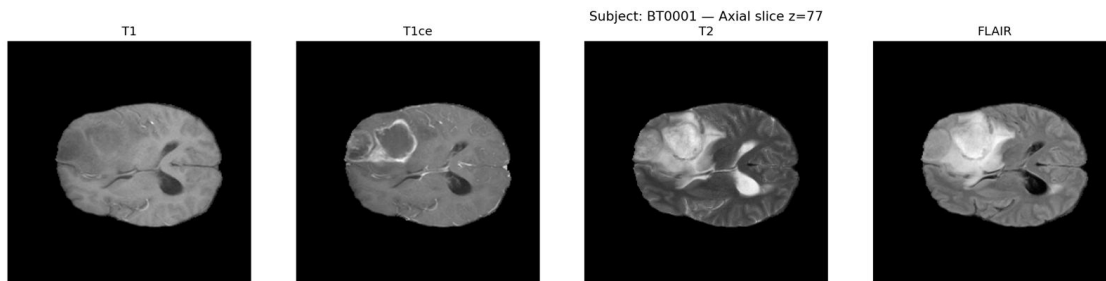


Fig. 1. Representative axial MRI slices across the four modalities

To address these challenges, we investigate a structured modality dropout strategy for MRI-based IDH gene mutation prediction. This approach exposes the model to clinically realistic modality availability scenarios, including missing T1CE, missing FLAIR, missing T2, only T1 and T2 available, and only T1 available. By simulating real-world modality-missing scenarios, the model learns modality-invariant features, which improves its robustness.

The proposed model was trained on the University of Texas Southwestern (UTSW) Glioma MRI dataset [14], comprising 622 subjects, including 176 IDH-mutated and 446 IDH-wildtype subjects. A 3D ResNet18 architecture [15] was trained and evaluated using five-fold stratified cross-validation. Performance was compared across three training strategies: a baseline model without dropout, a moderate dropout configuration, and a more aggressive dropout configuration. Performance was then evaluated across six

modality availability scenarios using metrics including AUC-ROC, balanced accuracy, sensitivity, specificity, calibration metrics, and bootstrap confidence intervals.

The main contributions of this study include a modality based structured dropout architecture for improving MRI based IDH mutation prediction in incomplete imaging scenarios, and evaluation of the model across six clinically realistic missing modality scenarios, allowing us to measure model performance as fewer imaging data is available. The study also includes comparison of standard training strategies against robustness-focused strategies using a large, multi-parametric glioma cohort, evaluated through cross validation.

2 Methodology

2.1 Dataset and Preprocessing

This study utilized the University of Texas Southwestern (UTSW) Glioma MRI Dataset [14], a multi-parametric retrospective MRI cohort comprising 622 subjects diagnosed with glioma. Of these, 176 subjects (28.3%) carried an IDH gene mutation (IDH-mutant), and 446 subjects (71.7%) were IDH-wildtype, reflecting the class imbalance characteristic of population-level glioma cohorts. Each subject had four MRI modalities available: T1-Weighted (T1), T1-Weighted with Contrast Enhancement (T1CE), T2-Weighted (T2), and Fluid-Attenuated Inversion Recovery (FLAIR). Cross-validation folds were constructed at the patient level, ensuring that all scans belonging to a given subject were assigned exclusively to a single fold and never split across training and validation partitions. Folds were stratified solely by IDH mutation status, preserving the 28.3%/71.7% class ratio across all five training and validation partitions. The data set contains scans acquired on multiple scanner platforms and field strengths, providing a naturalistic representation of clinical imaging variability.

To pre-process the MRI scans, each MRI volume was loaded in NIfTi format, and Z-score normalization was applied independently per modality, using the mean and standard deviation (SD) of non-zero voxels. The volumes were then resized to an isotropic resolution of $96 \times 96 \times 96$ by trilinear interpolation. The four normalized modalities were stacked along the channel dimension in a single four-channel tensor of shape $(4, 96, 96, 96)$. This resulting tensor was cast to a 32-bit floating point for model input.

The models were directly trained on the pre-processed, unaugmented tensors described above.

2.2 Model Architecture

A 3D ResNet-18 model was used as the classification backbone. ResNet-18 [15] was selected over other deeper ResNet or DenseNet variants due to the risk of overfitting due to the moderate cohort size (622 subjects). The residual connections preserve gradient flow through volumetric 3D input while keeping the parameter count and GPU memory footprint tractable throughout the training. The first convolutional layer was modified to accept input with four channels, corresponding to the four modalities. The model pro-

duced a 512-dimensional feature embedding at the penultimate layer. This signal was then passed through a fully connected classification layer with two output units and a softmax activation, trained with cross-entropy loss, for binary classification of IDH mutation status (mutated vs. wildtype). The remainder of the architecture followed the standard ResNet-18 structure, including residual connections and batch normalization at each block [16]. The modified backbone contains approximately 33 million trainable features, and weights were initialized using He (Kaiming) initialisation.

2.3 Structured Modality Dropout

To reflect how imaging data is collected in real-world scenarios and to make the model robust when some scans are missing, a structured modality dropout strategy was adopted, following the approach of Cheng et al. [12] and Zhou et al. [13]. During training, the model was exposed to six combinations of scans: all four modalities, missing T1CE, missing FLAIR, missing T2, only T1 and T2 available, and only T1 available.

Formally, let

$$X = [T_1, T_{1CE}, T_2, \text{FLAIR}]$$

denote the complete four-channel input. At each training iteration, one of the six modality-availability scenarios was sampled for a given training sample according to a fixed categorical distribution (Table 1), drawn independently and uniformly at random per sample, without weighting toward specific patients or folds. For a sampled scenario in which modality m was unavailable, the corresponding channel was replaced with a zero tensor of identical spatial dimensions, e.g., for missing T_{1CE} :

$$X' = [T_1, \mathbf{0}, T_2, \text{FLAIR}], \quad \mathbf{0} \in \mathbb{R}^{96 \times 96 \times 96}$$

Missing modalities were simulated by substituting the corresponding input channels with zero-filled volumes rather than mean imputation, Gaussian noise, or a learned placeholder token, since zero-filling introduces no additional learned parameters, avoids injecting dataset-specific statistical bias into the missing channel, and is a standard, easily reproducible convention for simulating missing-modality conditions in the multimodal medical imaging literature [12, 13]. This ensured the model did not prioritize any specific modality when other sequences were absent.

Table 1. Modality-availability scenario sampling probabilities per training configuration.

Scenario	Dropout-30	Dropout-50
All four modalities	40%	10%
Missing T1CE	15%	20%
Missing FLAIR	15%	20%
Missing T2	15%	20%
T1 + T2 only	10%	15%
T1 only	5%	15%

Three models were trained to understand how differently they react to robustness with dropout applied. The three models trained were the Baseline model, the Dropout-30

model, and the Dropout-50 model. The Baseline model was trained on all four modalities being present; no dropout action was applied to this model. The Dropout-30 model was exposed to all four modalities 40 percent of the time and distributed the remaining time across the five missing-modality scenarios, representing a moderate robustness strategy that still emphasizes complete-data training. The Dropout-50 model was trained focusing more on aggressive robustness, as it was exposed to all four modalities only 10 percent of the time, thus maximizing exposure to incomplete imaging conditions during training. The overall intent here was to stress-test the model’s robustness by treating incomplete imaging as the norm during training rather than an edge case.

2.4 Training Protocol

All models were trained using weighted binary cross-entropy loss to address class imbalance. Per-fold class weights were computed as the ratio of IDH-wildtype to IDH-mutant subjects within each training partition, assigning a proportionally higher penalty to misclassified IDH-mutant cases. Model parameters were optimized using the Adam optimizer [17] with default adam betas, $\beta_1 = 0.9$, $\beta_2 = 0.999$, a batch size of 24, an initial learning rate of 1×10^{-4} , and a weight decay of 1×10^{-4} , for a maximum of 100 epochs.

A ReduceLR0nPlateau learning rate scheduler was applied, halving the learning rate after 3 epochs without improvement in validation loss. Early stopping was employed, monitoring validation AUC and terminating training if no improvement was observed within a fixed patience window of 7 epochs. Gradient clipping with a maximum norm of 1.0 was applied at each update step to stabilize training. All experiments used a fixed random seed of 42 to ensure reproducibility.

All experiments were conducted on 6× NVIDIA DGX A100 GPU(s) within a Docker container, using PyTorch 2.6.0a0+df5bbc09d1.nv24.12, CUDA 12.6, and Python 3.12.3. Data handling and metric computation used NumPy 1.26.4 and scikit-learn 1.5.2 [18].

2.5 Experimental Design and Evaluation

Model performance was assessed using K -fold cross-validation [19], with $K = 5$ and folds stratified by IDH mutation status to preserve class proportions across training and validation partitions. Five-fold cross-validation was selected to balance reliable performance estimation against the computational cost of training three configurations across six modality-availability scenarios; with 622 subjects, five folds yield validation partitions of roughly 124 subjects each, sufficient to estimate AUC with reasonable precision while keeping the total number of trained models (3 configurations × 5 folds = 15) computationally tractable. A higher fold count (e.g., 10-fold) was considered but would have further reduced per-fold validation set size without a proportional gain in estimate stability, given the moderate cohort size. Each of the three training configurations (Baseline, Dropout-30, Dropout-50) was evaluated under all six modality-availability scenarios at inference time, yielding 18 condition-specific performance estimates per fold.

Primary performance metrics were calculated using scikit-learn [18] and included the area under the receiver operating characteristic curve (AUC-ROC), balanced accuracy, sensitivity, specificity, and the F1 score. A fixed classification threshold of 0.5 was applied

to the predicted probability of the IDH-mutant class to compute sensitivity, specificity, balanced accuracy, and F1 score.

Sensitivity (recall) indicates, of all actual IDH-mutant cases, how many were correctly identified:

$$\text{Sensitivity} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (1)$$

Specificity indicates, of all actual IDH-wildtype cases, how many were correctly identified:

$$\text{Specificity} = \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalsePositive}} \quad (2)$$

Balanced accuracy is the arithmetic mean of sensitivity and specificity, used specifically because of the class imbalance in this dataset ($\sim 28/72$ split). Regular accuracy would be misleading here, since a model that always predicts wildtype would achieve 72% accuracy.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (3)$$

F1 score was also reported, with all metrics expressed as mean $\pm$ SD across folds. Pooled out-of-fold (OOF) AUC was additionally computed by aggregating predicted probabilities across all validation folds, with 95% CI estimated via 1000-iteration stratified bootstrap resampling [20]. Probabilistic calibration was assessed using the Brier score [21]. Feature representations were examined through t-distributed stochastic neighbor embedding (t-SNE) [22] projections of 512-dimensional OOF embeddings extracted from the penultimate layer of each model under the all-four-modality scenario. Subgroup analyses were conducted by stratifying pooled OOF predictions by MRI scanner type and WHO tumor grade.

The two dropout configurations (30% and 50% labels) were selected to bracket a moderate and an aggressive robustness strategy rather than exhaustively sweep the dropout-rate space, given the computational cost of training and evaluating 15 full five-fold models across six modality scenarios each. A finer-grained sweep (e.g., 10/20/30/40/50%) is identified as future work.

3 Results

3.1 Classification Performance Across Modality Scenarios

Table 2 reports mean $\pm$ SD for all three models across AUC, balanced accuracy, sensitivity, specificity, and F1 score, for all six modality combinations, evaluated using five-fold stratified cross-validation [19].

Table 2. Classification performance (mean $\pm$ SD) across models and modality scenarios.

Model	Scenario	AUC	Balanced Acc.	Sensitivity	Specificity	F1
Baseline	All 4	0.823 $\pm$ 0.042	0.684 $\pm$ 0.052	0.584 $\pm$ 0.264	0.785 $\pm$ 0.196	0.527 $\pm$ 0.096
	No T1CE	0.792 $\pm$ 0.063	0.693 $\pm$ 0.059	0.635 $\pm$ 0.253	0.751 $\pm$ 0.144	0.534 $\pm$ 0.114
	No FLAIR	0.706 $\pm$ 0.044	0.574 $\pm$ 0.046	0.250 $\pm$ 0.178	0.897 $\pm$ 0.100	0.293 $\pm$ 0.164
	No T2	0.708 $\pm$ 0.033	0.591 $\pm$ 0.020	0.699 $\pm$ 0.259	0.483 $\pm$ 0.249	0.444 $\pm$ 0.074
	T1+T2 only	0.688 $\pm$ 0.041	0.593 $\pm$ 0.064	0.278 $\pm$ 0.196	0.908 $\pm$ 0.080	0.319 $\pm$ 0.206
	T1 only	0.529 $\pm$ 0.092	0.507 $\pm$ 0.026	0.434 $\pm$ 0.316	0.579 $\pm$ 0.322	0.301 $\pm$ 0.102
Dropout-30	All 4	0.774 $\pm$ 0.070	0.686 $\pm$ 0.082	0.601 $\pm$ 0.255	0.771 $\pm$ 0.208	0.538 $\pm$ 0.123
	No T1CE	0.763 $\pm$ 0.083	0.679 $\pm$ 0.095	0.533 $\pm$ 0.290	0.825 $\pm$ 0.153	0.483 $\pm$ 0.246
	No FLAIR	0.740 $\pm$ 0.064	0.680 $\pm$ 0.067	0.704 $\pm$ 0.107	0.657 $\pm$ 0.163	0.554 $\pm$ 0.074
	No T2	0.709 $\pm$ 0.053	0.631 $\pm$ 0.068	0.800 $\pm$ 0.174	0.461 $\pm$ 0.296	0.518 $\pm$ 0.045
	T1+T2 only	0.736 $\pm$ 0.074	0.659 $\pm$ 0.089	0.499 $\pm$ 0.255	0.818 $\pm$ 0.120	0.457 $\pm$ 0.235
	T1 only	0.580 $\pm$ 0.044	0.519 $\pm$ 0.036	0.657 $\pm$ 0.411	0.380 $\pm$ 0.439	0.359 $\pm$ 0.125
Dropout-50	All 4	0.708 $\pm$ 0.065	0.581 $\pm$ 0.087	0.243 $\pm$ 0.265	0.920 $\pm$ 0.092	0.251 $\pm$ 0.252
	No T1CE	0.706 $\pm$ 0.063	0.570 $\pm$ 0.067	0.182 $\pm$ 0.189	0.958 $\pm$ 0.058	0.232 $\pm$ 0.208
	No FLAIR	0.690 $\pm$ 0.056	0.582 $\pm$ 0.079	0.281 $\pm$ 0.305	0.884 $\pm$ 0.151	0.262 $\pm$ 0.239
	No T2	0.697 $\pm$ 0.056	0.571 $\pm$ 0.082	0.326 $\pm$ 0.386	0.816 $\pm$ 0.244	0.238 $\pm$ 0.253
	T1+T2 only	0.688 $\pm$ 0.052	0.564 $\pm$ 0.060	0.187 $\pm$ 0.182	0.942 $\pm$ 0.067	0.229 $\pm$ 0.204
	T1 only	0.634 $\pm$ 0.051	0.532 $\pm$ 0.036	0.337 $\pm$ 0.416	0.727 $\pm$ 0.351	0.198 $\pm$ 0.221

Fig. 2 shows the ROC curves across all six modality scenarios for the three models. When the complete four modality conditions were applied, the Baseline model performed the best, with AUC 0.823 ± 0.042 . This was followed by the Dropout-30 model (AUC 0.774 ± 0.070) and the Dropout 50 model (AUC 0.708 ± 0.065). The findings make it clear that when all MRI sequences are available, the model trained on complete data has an edge over models exposed to incomplete data. However, this advantage diminishes quickly when the number of available modalities decreases.

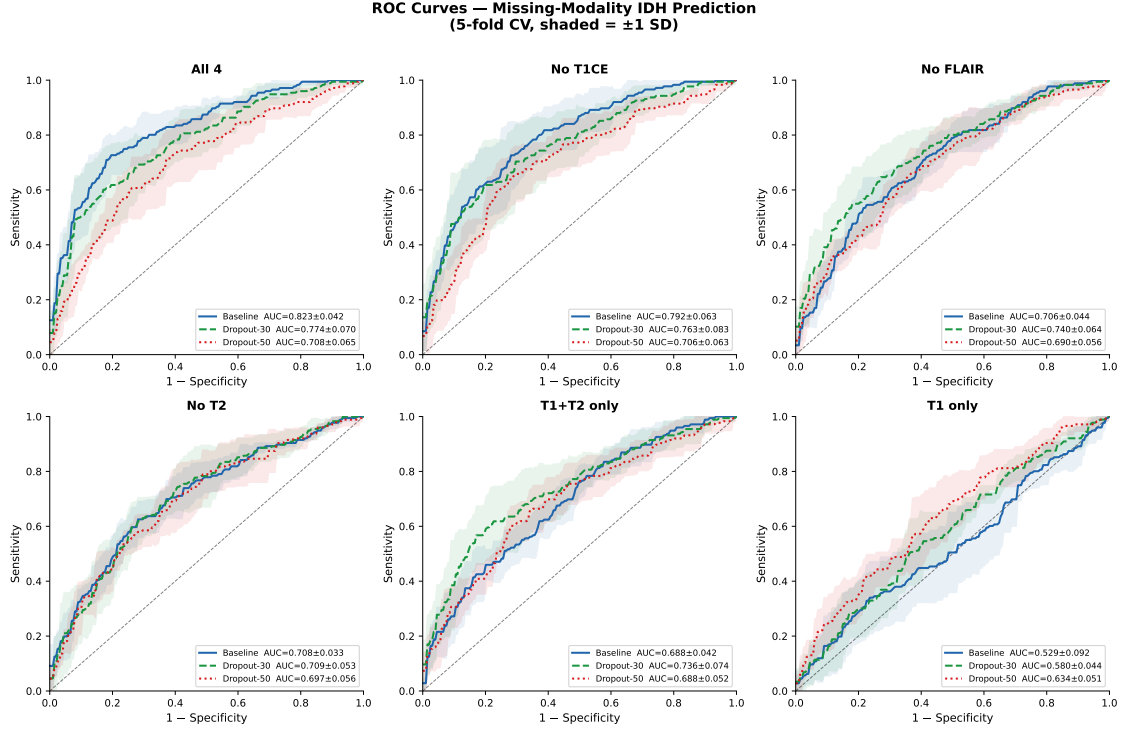


Fig. 2. ROC curves across all six modality scenarios comparing the Baseline, Dropout-30, and Dropout-50 models

When T1CE modality was missing, the Baseline model showed a moderate decline in performance (AUC 0.792 ± 0.063). A proportionate decline was also seen in the Dropout-30 (AUC 0.763 ± 0.083) and Dropout-50 models (AUC 0.706 ± 0.063). This finding suggests that the available modalities - T1, T2 and FLAIR - carry significant and overlapping information.

In contrast, when the FLAIR modality was made unavailable, the Baseline model (AUC 0.706 ± 0.044) suffered a worse decline as compared to Dropout-30 (AUC 0.740 ± 0.064) and Dropout-50 (AUC 0.690 ± 0.056) models. The Dropout-30 model performed the best in this scenario, outperforming the Baseline by 3.4 AUC points. The Dropout-30 model also performed better than the baseline under the T1+T2 condition as well.

The most challenging scenario was the single modality T1 input, where the performance of Baseline suffered as performed close to chance (AUC 0.529 ± 0.092). The Dropout-30 model was slightly more resilient (AUC 0.580 ± 0.044), while the Dropout-50 model performed best in this scenario (AUC 0.634 ± 0.051). This shows that aggressive dropout training prepared the model to extract useful features from very limited data. The findings point to a fundamental tradeoff: aggressive dropout training allows the model to perform better when data is severely limited, but this comes at the cost of peak performance when complete data is available. Fig. 3 summarizes this trend, plotting mean AUC as a function of the number of available modalities for each model.

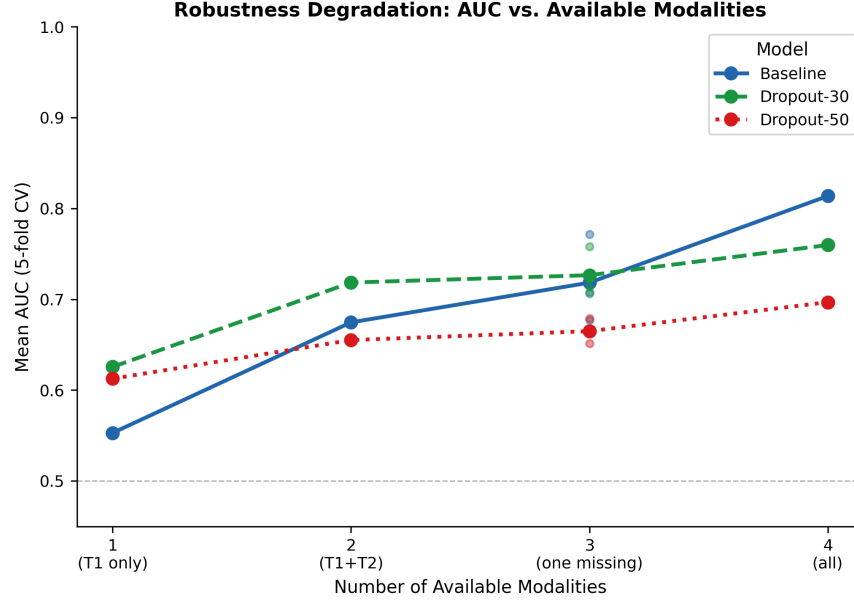


Fig. 3. Mean AUC versus number of available MRI modalities for the Baseline, Dropout-30, and Dropout-50 models

The Baseline model showed low sensitivity when FLAIR was missing (0.250 ± 0.178) and under the T1+T2-only scenario (0.278 ± 0.196), with correspondingly high specificity (0.897 and 0.908, respectively), indicating that when key modalities were removed the model leaned heavily toward predicting the majority class rather than effectively discriminating between classes. Dropout-30, by contrast, struck a much better balance between sensitivity and specificity in these same scenarios – 0.704 ± 0.107 for missing FLAIR and 0.499 ± 0.255 for T1+T2-only – suggesting more effective feature learning across conditions.

3.2 Pooled Out-of-Fold AUC and Calibration

Table 3 shows the pooled out-of-fold (OOF) AUC computed across all validation subjects, with 1000-iteration stratified bootstrap 95% CI [20], along with Brier scores for each model.

Table 3. Pooled out-of-fold AUC (95% CI) and Brier scores by model.

Model	Pooled AUC (95% CI)	Brier Score
Baseline	0.772 (0.729–0.812)	0.217
Dropout-30	0.744 (0.699–0.790)	0.198
Dropout-50	0.6369 (0.5860, 0.6854)	0.1996

The Baseline model achieved the highest pooled AUC of 0.772 (95% CI: 0.729–0.812), followed by Dropout-30 at AUC 0.744 (95% CI: 0.699–0.790). The confidence intervals of Baseline and Dropout-30 overlap partially, suggesting a moderate performance difference

between the two models, whereas Dropout-50 demonstrated a more substantial reduction in performance.

AUC measures the discriminative ability of a model, whereas the Brier score [21] measures calibration – the mean squared difference between predicted probability and actual outcome (0 or 1). A lower Brier score indicates better agreement between predicted probabilities and observed outcomes. Dropout-30 achieved the lowest Brier score of 0.198, compared to 0.217 for Baseline and 0.200 for Dropout-50. This higher calibration for Dropout-30 suggests that a moderate dropout training strategy produces better-calibrated predictions while preserving discriminative performance.

3.3 Feature Embedding Visualization

Fig. 4 shows t-SNE projections of the OOF embeddings extracted from each model under the all-four-modality scenario.

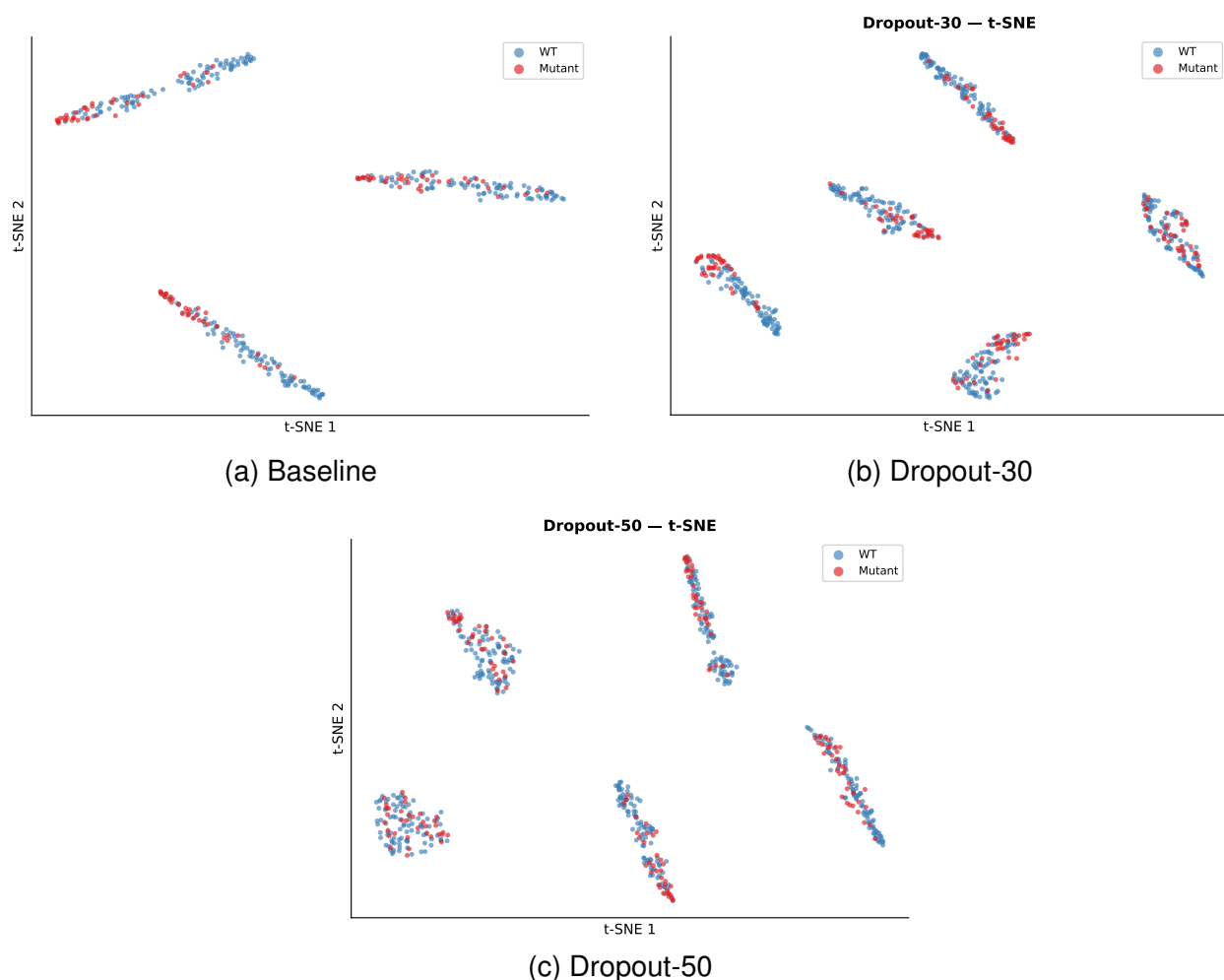


Fig. 4. t-SNE plots [22] of 512-dimensional OOF embeddings from the penultimate layer of each model under the all-four-modality scenario. Blue = IDH WT; red = IDH-mutant

The Baseline model produced elongated, mostly separated clusters; IDH-mutant (red)

and IDH-wildtype (blue) patients formed largely linear structures through the embedding space. This suggests the Baseline network produced feature representations based almost entirely on all-four-modality inputs, providing moderate class separation. The Dropout-30 embeddings were less cohesive and more distributed, with some inter-class mixing, suggesting the network learned feature representations distributed over several sub-spaces rather than a single primary discriminative dimension. The Dropout-50 embeddings showed the greatest dispersion and overlap between IDH-mutant and IDH-wildtype groups, consistent with its lower discriminative capability under full-modality conditions. These results indicate that structured modality dropout changes the topology of the learned feature space, allowing for more distributed representations while reducing class compactness when all modalities are present.

3.4 Subgroup Analysis by Scanner and Tumor Grade

Fig. 5 presents AUC stratified by MRI scanner type and by WHO tumor grade for the three models.

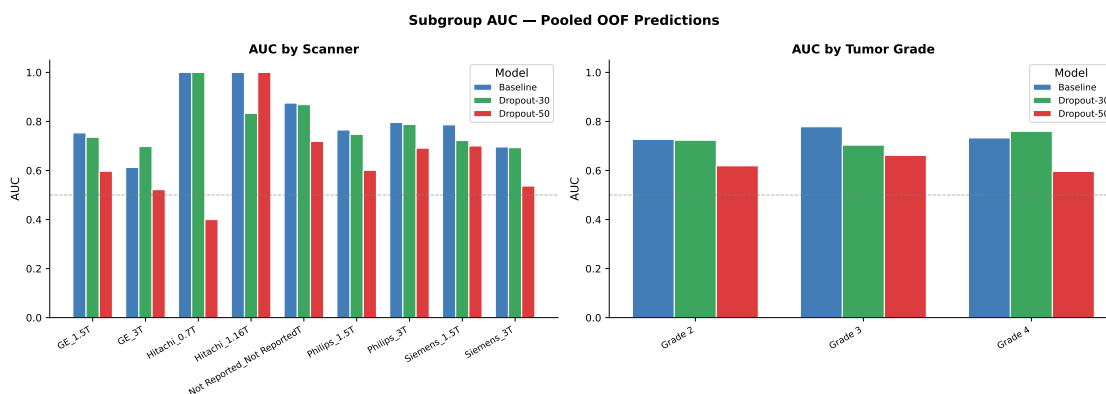


Fig. 5. AUC by MRI scanner type (left) and by tumor grade (right) for the Baseline, Dropout-30, and Dropout-50 models, based on pooled out-of-fold predictions

Scanner and tumor grade subgroup analyses were conducted on pooled OOF predictions across platforms consistent with those reported in multi-institutional glioma imaging benchmarks [23–28].

Scanner-stratified results showed generally consistent performance for the Baseline and Dropout-30 models across major platforms. Both models achieved $AUC < 0.70$ for GE 1.5T, Philips 1.5T, Philips 3T, Siemens 1.5T, and Siemens 3T scanners, indicating reasonable cross-scanner generalizability. Ideal AUC values of 1.000 were observed for both Baseline and Dropout-30 on the Hitachi 0.7T and Hitachi 1.16T subgroups; however, these subgroups have small sample sizes, and these values should be interpreted with caution. Dropout-50 showed markedly degraded performance on several scanner platforms, including GE 3T (0.522), Hitachi 0.7T (0.400), and Siemens 3T (0.537), suggesting that overly aggressive modality dropout during training may compromise the network's ability to generalize across scanner-specific imaging characteristics.

Tumor grade subgroup analysis showed that all three models maintained AUC values in the range of 0.60–0.78 across WHO Grades 2, 3, and 4. The Baseline model performed best on Grade 3 tumors (AUC 0.778), while Dropout-30 achieved the highest AUC among all models on Grade 4 gliomas (0.760 vs. 0.733 for Baseline). Grade 2 tumors, which are predominantly IDH-mutant, showed the most consistent performance across models (0.619–0.727), potentially reflecting the strong association between lower-grade histology and IDH mutation status captured by all models. Dropout-50 consistently underperformed relative to the other two models across all grade subgroups, with AUC values ranging from 0.596 to 0.662.

In general, the subgroup analyses suggest that Dropout-30 offers the most favorable balance between scanner generalizability and discriminative performance specific to the grade, making it the most clinically viable candidate among the three training strategies evaluated.

4 Conclusion

In real-world clinical environments, complete imaging data may or may not be available. This study explores whether a structured modality dropout training strategy could improve the robustness of IDH mutation prediction using MRI scans. The IDH mutation status is an important marker in brain tumor analysis, indicating factors such as favorability of prognosis and response to treatment. We developed three training strategies, a Baseline model, a moderate Dropout-30 model, and an aggressive Dropout-50 model. All three models were evaluated on six real world modality missing scenarios. The models were validated through five-fold stratified cross validation on a cohort of 622 subjects. The findings show that when all four MRI modalities are available, the Baseline model performed the best, but performance declined when the number of missing modalities increased. The Dropout-30 model sacrificed peak performance, but its performance was relatively stable across all the modality combinations. It was also able to achieve the best Brier score among all the models. The Dropout-50 model performed the best in extreme single modality scenarios, but at the cost of low overall performance and cross-scanner generalizability. The results show that the model choice depends on the expected completeness of the imaging data. The Dropout-30 model is the most balanced and clinically viable strategy, taken together with scanner and tumor grade stratified subgroup analysis. The limitations of this study include the lack of external validation, moderate class imbalance, and the lack of significance of pairwise statistical comparisons due to limited statistical power of five-fold cross validation. The scope of future work includes testing the findings on large, multi-institutional cohorts, and applying the modality dropout framework to other missing data combinations and its applicability to other molecular markers beyond IDH mutation status. Overall, the study demonstrates a practical strategy for developing MRI based deep learning models that are effective in real world clinical settings.

Declarations

Funding Statement

This research did not receive external funding.

Ethics Statement

This study used the retrospective, de-identified University of Texas Southwestern (UTSW) Glioma MRI dataset [14]. The original data collection was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of the University of Texas Southwestern Medical Center (STU-2020-1184). As this study involved only secondary analysis of retrospective, de-identified imaging data obtained from a public repository (The Cancer Imaging Archive), no additional institutional review was required for the present analysis. No new human or animal data were collected as part of this study.

Consent to Participate

Consent to Participate declaration: not applicable.

Consent to Publish

Consent to Publish declaration: not applicable.

Author Contributions

M.M. and P.G. contributed equally to conducting the experiments, developing the code, and writing the manuscript. S.P.S. supervised the project, guided the study design, contributed to the code, and edited and reviewed the manuscript.

Data availability

The MRI data used in this study are derived from the University of Texas Southwestern (UTSW) Glioma MRI dataset [14], available at <https://www.cancerimagingarchive.net/collection/utsw-glioma/>.

References

- [1] S. Valvi, M. Fouladi, M. J. Fisher, and N. G. Gottardo. Idh mutant high-grade gliomas. *Frontiers in Molecular Neuroscience*, 18, 2025. doi: 10.3389/fnmol.2025.1662414.
- [2] B. T. Whitfield and J. T. Huse. Classification of adult-type diffuse gliomas: Impact of the world health organization 2021 update. *Brain Pathology*, 32(4):e13062, 2022. doi: 10.1111/bpa.13062.

- [3] Y. W. Park, K. Han, S. S. Ahn, S. Bae, Y. S. Choi, J. H. Chang, and S. K. Lee. The 2021 who classification for gliomas and implications on imaging diagnosis: Part 1—key points of the fifth edition and summary of imaging findings on adult-type diffuse gliomas. *Journal of Magnetic Resonance Imaging*, 58(3):677–689, 2023. doi: 10.1002/jmri.28743.
- [4] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffiatti, A. Von Deimling, and D. W. Ellison. The 2021 who classification of tumors of the central nervous system: A summary. *Neuro-Oncology*, 23(8):1231–1251, 2021. doi: 10.1093/neuonc/noab106.
- [5] X. Chen, J. Lei, S. Wang, J. Zhang, and L. Gou. Diagnostic accuracy of a machine learning-based radiomics approach of mr in predicting idh mutations in glioma patients: A systematic review and meta-analysis. *Frontiers in Oncology*, 14:1409760, 2024. doi: 10.3389/fonc.2024.1409760.
- [6] C. Y. C. Chung and L. E. Pigott. Predicting idh and atrx mutations in gliomas from radiomic features with machine learning: A systematic review and meta-analysis. *Frontiers in Radiology*, 4:1493824, 2024. doi: 10.3389/fradi.2024.1493824.
- [7] D. N. G. Kasap, N. G. Nacul Mora, D. A. Blömer, B. H. Akkurt, W. L. Heindel, M. Manuil, and M. Musigmann. Comparison of mri sequences to predict idh mutation status in gliomas using radiomics-based machine learning. *Biomedicines*, 12(4):725, 2024. doi: 10.3390/biomedicines12040725.
- [8] S. Kino, M. Kanamori, Y. Shimoda, K. Niizuma, H. Endo, and Y. Matsuura. Distinguishing idh mutation status in gliomas using ftir-atr spectra of peripheral blood plasma indicating clear traces of protein amyloid aggregation. *BMC Cancer*, 24(1), 2024. doi: 10.1186/s12885-024-11970-y.
- [9] S. Chakrabarty, P. LaMontagne, J. Shimony, D. Marcus, and A. Sotiras. Non-invasive classification of idh mutation status of gliomas from multi-modal mri using a 3d convolutional neural network. In *Proceedings of SPIE*, volume 12465, page 124650W, 2023. doi: 10.1117/12.2651391.
- [10] C. G. B. Yogananda, B. C. Wagner, N. C. D. Truong, J. M. Holcomb, D. D. Reddy, N. Saadat, K. J. Hatanpaa, T. R. Patel, B. Fei, M. D. Lee, R. Jain, R. J. Bruce, M. C. Pinho, A. J. Madhuranthakam, and J. A. Maldjian. Mri-based deep learning method for classification of idh mutation status. *Bioengineering*, 10(9):1045, 2023. doi: 10.3390/bioengineering10091045.
- [11] M. Griessmair, C. Delbridge, J. Ziegenfeuter, D. Bernhardt, J. Gempt, F. Schmidt-Graf, B. Wiestler, and M. C. Metz. Imaging the who 2021 brain tumor classification: Fully automated analysis of imaging features of newly diagnosed gliomas. *Cancers*, 15(8):2355, 2023. doi: 10.3390/cancers15082355.

- [12] R. Cheng, Z. Sun, Y. Zhang, and C. Li. Robust divergence learning for missing modality segmentation. *arXiv preprint arXiv:2411.08305*, 2024. URL <http://arxiv.org/abs/2411.08305>.
- [13] T. Zhou, S. Canu, P. Vera, and S. Ruan. Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing mr modalities. *Neurocomputing*, 466:102–112, 2021. doi: 10.1016/j.neucom.2021.08.081.
- [14] D. Reddy, N. Saadat, J. Holcomb, B. Wagner, N. Truong, J. Bowerman, K. Hatanpaa, T. Patel, M. Pinho, F. Yu, K. Zhang, S. Lodhi, A. Madhuranthakam, C. G. Bangalore Yogananda, and J. Maldjian. The university of texas southwestern glioma MRI dataset with molecular marker characterization and segmentations (UTSW-glioma), 2026. URL <https://doi.org/10.7937/DFAE-1B86>. Version 1.
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90.
- [16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, volume 37, pages 448–456. JMLR.org, 2015. URL <http://proceedings.mlr.press/v37/ioffe15.html>.
- [17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015. URL <https://arxiv.org/abs/1412.6980>.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [19] G. Varoquaux and O. Colliot. Evaluating machine learning models and their diagnostic value. In *Machine Learning for Brain Disorders*. Springer, 2023. doi: 10.1007/978-1-0716-3195-9_18.
- [20] J. Pang, W. Ju, M. Welch, P. Gauger, P. Liu, Q. Zhang, and C. Wang. Nonparametric bootstrap methods for interval estimation of the area under the roc curve with correlated diagnostic test data. *Frontiers in Veterinary Science*, 10, 2023. doi: 10.3389/fvets.2023.1274786.
- [21] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
- [22] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

- [23] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, and K. Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.
- [24] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4: 170117, 2017. doi: 10.1038/sdata.2017.117.
- [25] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, and S. Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.
- [26] F. Kofler, D. Waldmannstetter, F. Steinbauer, S. Bakas, B. H. Menze, B. Wiestler, and I. Ezhov. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn). *arXiv preprint arXiv:2305.09011*, 2023.
- [27] D. LaBella, M. Adewole, M. Alonso-Basanta, R. Bhalerao, R. R. Bhatt, U. Baid, and S. Bakas. The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. *arXiv preprint arXiv:2305.07642*, 2023.
- [28] A. W. Moawad, R. Jain, R. R. Bhatt, E. Calabrese, E. Colak, A. B. Mahmoud, and S. Bakas. The brain tumor segmentation (brats) challenge 2023: Brain mr image segmentation for central nervous system metastases. *arXiv preprint arXiv:2306.00838*, 2023.