Executive Summary: Brain Tumor Diagnosis

Chris Ewasiuk, Jacob Johnson, Wenwen Li, Suo-Jun Tan

Project Github

https://github.com/Erdos-Projects/fall-2025-brain-tumor-diagnosis/

Problems

- Accurate diagnosis of brain tumors relies heavily on expert radiologists manually reviewing MRI scans; a process that is both time-consuming and subject to human variability. Moreover, constructing labeled datasets for training diagnostic models requires precise manual segmentation of tumor regions, which further limits scalability. This motivates the development of automated methods that can rapidly and reliably identify the presence of tumors and generate accurate 3D masks of their extent within the brain.
- Glioblastoma (GBM; WHO CNS Grade 4) is the most aggressive primary brain tumor. Early and accurate differentiation between GBM and other gliomas (OG; WHO CNS Grades 2-3) directly influences treatment planning, clinical trial eligibility, and patient counseling. Manual MRI interpretation is time-consuming and subject to inter-observer variability, particularly in borderline cases.
- IDH mutation plays an important role in the prognosis of diffuse gliomas. Generally, IDH-mutant type has a better prognosis and more treatment options compared to IDH-wildtype. Contrast-enhanced T1-weighted images are often cited as the most useful among conventional sequences, but they are still less reliable than a direct molecular test, which can be invasive and costly.

Objective

- Develop a 3D Convolutional Neural Network (CNN) capable of performing tumor versus no-tumor classification while rapidly producing voxel-level 3D segmentation masks that replicate expert precision.
- Develop binary classifier by fine-tuning a pre-trained 18-layer Residual Network using individual slices of the 3D MRI along all three coordinate axes to classify images in tumor/no-tumor classes.
- Develop a convolutional neural network (CNN) to classify preoperative MRI scans as Glioblastoma (GBM) vs. Other Gliomas (OG) using MRI data and clinical metadata.
- Develop a simple 2D Convolutional Neural Network (CNN) to identify IDH mutation status on MRI images.

Data Sources

- Public UCSF-PDGM dataset: preoperative MRI scans and clinical metadata of preoperative diffuse gliomas. https://www.cancerimagingarchive.net/collection/ucsf-pdgm/
- Modalities: ADC, FLAIR, T1, T1-Gd, T2.
- Labels: WHO grade[2], IDH mutation status, 1p/19q co-deletion.
- After preprocessing (removing follow–ups): **495 patients**.
- Random cross-sectional slices taken along each of the three coordinate axes were taken from each 3D MRI volume are saved as grayscale .png images for use in the 2D binary classifier.
  - 4 random slices were taken along each coordinate axis in regions containing tumors as well as regions within the brain volume containing no tumor, using the brain segmentation and tumor segmentation maps to determine regions containing brain matter and tumor masses.
  - A minimum cross-sectional area of 30% of the maximum cross-sectional area of the MRI slices along the respective coordinate axis to guarantee meaningfully sized cross-section of the brain for each image.
  - If fewer than 4 qualifying slices exist along any axis, then all qualifying slices are saved as images.
  - Total images saved containing tumors for each modality: 5940
  - Total images saved containing no tumors for each modality: 5931
  - All images are saved in a google drive here.

Imaging input consists of one representative slice per modality (2D). Metadata features (age, sex, IDH status, 1p/19q status) are integrated to align with the WHO 2021 molecular classification standards.

Key Performance Indicators (KPIs)

- Accuracy and Balanced Accuracy
- Sensitivity (recall) for GBM detection (clinical safety metric)
- Precision, F1 Score, AUROC score, Dice metric, and confusion matrix
- Calibration curves assessing probability reliability

Results

2D Tumor / No-Tumor Segmentation

Metric	Performance (across folds)
Threshold where relevant	> 0.5
Average Precision	$(0.89, 0.91, 0.91, 0.91)$
ROC AUC	$(0.87, 0.89, 0.89, 0.89)$
Accuracy	$(0.79, 0.76, 0.82, 0.81)$
Test AUC	0.89
Test Average Precision	0.91
Test Accuracy	0.80

3D Tumor / No-Tumor Segmentation

Metric	Performance
Mean Dice	∼0.83
Precision	∼ 86%
Recall	∼ 82%
Patients with Dice > 0.8	20 / 25
Low-Quality Scans (Dice < 0.3)	3 / 25
Overall Performance	Stable across grades 2–4

For the 3D tumor/no-tumor segmentation, a Dice metric approaching 1 indicates great overlap between the tumor truth mask and the model's predicted mask. These values are consistent with known performance bands for slice-based MRI glioma classification on limited datasets, indicating the model captures biologically relevant image features.

For the 2D tumor/no-tumor binary classifier, several neural network architectures were tested with varying performance. A 2D CNN designed similar to the CNN described in this paper[3] was iterated upon early in the modeling phase. The model suffered from slow convergence and poor performance, leading to a shift in the model architecture from a simple CNN to a pre-trained 18-layer residual network outlined in this paper [1]. The inputs to the final model architecture were chosen from the images constructed from only the FLAIR MRI measurements.

The fully connected layer was converted from a 1000 category output to a single category output for binary classification. During initial testing, it was found that updating all of the model parameters lead to egregious overfitting. All of the residual block layer parameters were frozen during fine-tuning, leaving only the fully-connected layer parameters open for training. Dropout regularization was implemented in the fully connected layer to further reduce overfitting to the training set.

Data were split between training (80%) and test (20%) sets. In the interest of time, a 4-fold cross-validation scheme was implemented on the training set to evaluate the performance of the models. Validation set accuracy was used as the metric to take the "best" model parameters over a small fine-tuning period of 5 epochs per fold. Models did not show significant overfitting during this short training period once dropout regularization was implemented. Model performance on the validation sets were similar across folds. The model from the third split was chosen to check the performance against the test set.

Glioblastoma/Other Gliomas Classifier

Metric	Performance
Training Accuracy	∼91%
Balanced Validation Accuracy	∼86%
Balanced Test Accuracy	∼81%
Validation GBM Recall	∼86%
Validation Other Glioma Recall	∼85%
Test GBM Recall	∼81%
Test Other Glioma Recall	∼80%
Test F1 Score	∼0.87
Test AUROC Score	∼0.79

IDH Mutation Classifier

Metric	Performance
Training Accuracy	∼90%
Test Accuracy	∼85%
Balanced Test Accuracy	∼77%
Test IDH mutant Recall	∼62%
Test IDH wildtype Precision	∼90%
Test IDH mutant Recall	∼62%
Test IDH wildtype Precision	∼68%
Test F1 Score	∼0.78
Test AUROC Score	∼0.81

For both the GBM/OG and the IDH mutation classifiers, several CNN architectures were tested to assess the feasibility of MRI-based prediction. Initial lightweight 2D CNNs showed unstable training and limited generalization, so we adopted custom 2D CNNs using five MRI modalities (ADC, FLAIR, T1, T1c, T2) with metadata fusion. The GBM/OG model used age, sex, IDH status, and 1p/19q status, while the IDH model used age and sex only. Data were stratified at the patient level and split into training (60%), validation (20%), and test (20%) sets to avoid patient-level leakage.

To address class imbalance (GBM: 396 vs. OG: 99), in the GBM/OG classifier, we oversampled the minority class only within the training data by extracting additional adjacent slices centered on the slice with the maximal tumor cross-section for OG cases. Validation and test sets remained imbalanced to preserve real-world class distribution. The model was selected based on validation accuracy and balanced accuracy over a 20-epoch fine-tuning window, and it demonstrates consistent improvement across training epochs without significant overfitting. The model also achieved strong performance on the test set. In particular, this binary classifier maintains balanced detection across GBM and OG, with OG recall of 80% and GBM recall of 81%, which is clinically important, as OG are rarer and more heterogeneous than GBM. Prior studies have shown that glioma-grade prediction models tend to show reduced performance on OG when class balance is not carefully addressed [4]. These results indicate that our GBM/OG classifier captures biologically meaningful features rather than over-fitting to class imbalance.

For the IDH mutation classifier, model selection was based on validation accuracy and validation loss over a 20-epoch fine-tuning window. The classifier performs better on the IDH wildtype gliomas (with 92% accuracy) while only identifying about 62% IDH mutant type gliomas in the test set correctly. While the precision gap between the two classes is narrower (90%

for wildtype and 68% for mutant), the PR curve for mutant type is visually flatter and more rapidly decreasing as recall increases. This indicates that further quality data collection could be potential helpful for the model.

Limitations

Tumor vs. No-Tumor Classification

- Dependence on clear, standardized scans
- Voxel-level imbalance (far more non-tumor voxels than tumorous)
- Tumor only dataset may introduce modeling bias and may not generalize to healthy scans
- Limited training to FLAIR,T1,T2 modalities, may not generalize to other scans
- MRI data are already segmented and contain no tissue other than brain matter, may not generalize to MRI data containing the skull and non-brain tissue.
- Sliding window used to scan full brain may impact segmentation analysis or image recombination

GBM/OG and the IDH Mutation Classification

- Limited dataset size (~495 patients)
- Single-slice input underrepresents full tumor morphology
- Single-institution dataset (generalizability constraints)
- Some label uncertainty due to evolving WHO classification criteria
- Class imbalance may bias prediction despite oversampling strategies
- Both models are binary and do not separate Grade 2 from Grade 3 gliomas, limiting fine-grained grading capability
- Mild calibration overconfidence observed

Future Directions

- Improve feature-based classifiers by gathering more quality data (especially for lower-grade tumors). Including other molecular markers could help distinguish between grades 2 and 3 tumors.
- Incorporate both models into a single model capable of performing both tumor versus no-tumor classification and tumor type identification on images determined to have a tumor.
- Expand from binary (GBM vs. OG or IDH wild vs. mutant) to full WHO grade classification (II / III / IV):
  - Handle grade II vs III distinction, which remains clinically subtle.
  - Incorporate additional MRI modalities (e.g., DSC perfusion, advanced diffusion, MR spectroscopy, SWI) because each provides unique biological information — improving glioma grading and molecular biomarker prediction.
  - Explore 3D CNNs + volumetric tumor masks.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1512.03385.

[2] D. N. Louis et al., "The 2021 who classification of tumors of the central nervous system: A summary", *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021. DOI: 10.1093/neuonc/noab106.

[3] S. Saeedi, S. Rezayi, H. Keshavarz, and S. R Niakan Kalhori, "MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques", en, *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 16, Jan. 2023.

[4] R. Sánchez-Marqués, V. García, and J. S. Sánchez, "A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance TCGA data", *Scientific Reports*, vol. 14, p. 17 195, 2024. DOI: 10.1038/s41598-024-68291-0.