Automated brain tumor response assessment from longitudinal multiparametric MRI data using Swin UNETR and a radiomics based classifier

Satyajit Maurya[1]*, Ewunate Assaye Kassaw[1]*, Mohammad Tufail Sheikh[1], Amit Mehndiratta[1,2,3,4], Anup Singh[1,2,3]

[1] Centre for Biomedical Engineering, Indian Institute of Technology Delhi, India
[2] Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, India
[3] Department of Biomedical Engineering, All India Institute of Medical Science, Delhi, India
[4] Faculty of Medicine and Health, University of New South Wales (UNSW), Sydney, Australia

anupsm@iitd.ac.in

(*Authors share equal contribution)

Abstract. Accurate and consistent response assessment is essential for guiding clinical decisions and optimizing treatment strategies in brain tumor patients. However, current methods for treatment response evaluation rely heavily on manual assessment of Response Assessment in Neuro-Oncology (RANO) criteria, which is time-consuming and prone to inter-observer variability. To address these limitations, we developed a fully automated pipeline combining segmentation and classification models to assess brain tumor response. Initially, Swin UNETR and U-Net models were trained on the BraTS dataset to automatically segment the whole tumor (WT) and enhancing tumor (ET) masks from FLAIR and WT masked T1c MRI sequences, respectively. Following segmentation from the best performing Swin UNETR model, shape-based and first-order radiomic features were extracted from the longitudinal LUMIERE dataset. A classification model utilizing TabM was developed for classifying the tumor treatment response into one of Complete Response (CR), Partial Response (PR), Stable Disease (SD), or Progressive Disease (PD) classes. Median Dice scores of 0.8811 and 0.8754 were obtained for the WT and ET using Swin UNETR models, respectively on the BraTS dataset. Using the extracted radiomic features, the TabM classifier achieved an average 5-fold cross-validation balanced accuracy of 0.6415 on the LUMIERE dataset. These results demonstrate the feasibility of automatically assessing brain tumor treatment response using longitudinal FLAIR and T1c MRI scans.

Keywords: Glioblastoma, Response Assessment in Neuro-Oncology (RANO) criteria, Swin UNETR, TabM.

1 Introduction

Gliomas and Meningiomas are amongst the most prevalent brain tumors. Gliomas account for almost 80% of all malignant brain tumors [1]. These are associated with high

rates of morbidity and mortality [2]. Glioblastoma (GBM) is the most aggressive form of glioma having a median overall survival of only 12-18 months and a five-year survival rate of around 5% [3]. The current standard of care for GBM involves maximal safe resection followed by adjuvant radiotherapy and chemotherapy, typically with temozolomide [4]. Despite advances in brain tumor care, therapeutic efficacy remains limited by the immunosuppressive tumor microenvironment and the restrictive blood-brain barrier, which impedes drug delivery and immune cell infiltration.

A critical and crucial aspect of managing brain tumor patients is the accurate and consistent assessment of their response to treatment. This evaluation guides crucial clinical decisions, such as continuing, modifying, or discontinuing therapies. To standardize the evaluation of treatment response, a set of criteria is essential, particularly in managing high-grade gliomas, where accurate and consistent assessment guides critical clinical decisions. The Macdonald criteria, introduced in 1990 relied on the bidimensional measurements of contrast-enhancing tumor regions [5]. While an important first step, this criterion proved to have significant limitations. For instance, they did not account for non-enhancing tumor components visible on T2-weighted and FLAIR MRI sequences. To address these shortcomings, the international Response Assessment in Neuro-Oncology (RANO) working group was established. The RANO criteria also incorporated the evaluation of non-enhancing T2/FLAIR signal abnormalities and integrated corticosteroid dosage and clinical status to provide a more holistic assessment. While RANO criteria provides a structured framework, its reliance on subjective imaging assessments and T2/FLAIR interpretation underscores the need for advanced criteria to standardize response classification.

The reliance on manual, bidimensional measurements is laborious, susceptible to inter as well as intra-observer variability, and can be inaccurate for the irregularly shaped tumors often seen in gliomas. This has led to a growing recognition of the advantages of volumetric analysis, which provides a more sensitive and reproducible method for quantifying tumor burden. Studies have shown that volumetric measurements can offer improved sensitivity in detecting subtle changes in tumor size over time [6]. Recognizing this, the RANO 2.0 criteria now formally includes volumetric measurements as an optional assessment method. Tumor response can be classified into categories such as complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD) based on a combination of imaging metrics, clinical status, and corticosteroid dosing. For high-grade gliomas, CR requires complete disappearance of all measurable contrast-enhancing disease, demonstrated on post-radiotherapy MRI with no new lesions and maintained for at least four weeks off corticosteroid therapy. PR is defined by at least a 50% reduction in the sum of products of perpendicular diameters of contrast enhancing lesions or a corresponding volumetric reduction (commonly ≥65%). This is accompanied by the additional condition that there is no increase in non-measurable disease, no appearance of new lesions, and no requirement for increased corticosteroid dosing. PD is characterized by a ≥25% increase in the product of diameters of enhancing lesions or a ≥40% increase in lesion volume and/or the emergence of new lesions. SD is designated when the changes in imaging findings do not meet the thresholds for CR, PR, or PD.

A key limitation of conventional RANO assessment is its dependence on manual lesion segmentation and bi-dimensional measurements. To overcome this challenge, some AI-driven approaches have been developed that demonstrate strong concordance with manual RANO classification while automating the measurement of contrast-enhancing lesions [7] [8]. AI-based algorithms, particularly those using deep learning, have demonstrated remarkable potential to automate and standardize the response assessment process [9] [10]. Some studies have aimed to develop end-to-end pipelines for assessing the tumor response [6]. However, these methods are limited either by small dataset sizes and poor classification accuracy, or by not classifying treatment response according to the RANO criteria into four classes [11].

The Swin UNETR [12] architecture emerges as a particularly promising solution, combining Swin Transformers [13] with U-Net [14] models to simultaneously capture local tissue details and global contextual relationships in 3D medical images. Following segmentation, radiomics analysis enables comprehensive tumor characterization through high-throughput extraction of quantitative features. These features encode tumor heterogeneity and tumor microenvironmental properties beyond human visual perception. When coupled with deep learning classifiers, this combined approach (Swin UNETR + radiomics) can offer a robust, data-driven framework for treatment response prediction that addresses the limitations of conventional methods while maintaining clinical interpretability.

This study, therefore, aims to develop and validate a fully automated pipeline for brain tumor response assessment from longitudinal multiparametric MRI data to overcome the subjectivity and labor-intensive nature of current RANO assessment methods utilizing the LUMIERE dataset [15] as part of the 2025 BraTS Brain Tumor Progression Challenge.

2 Materials and Methods

2.1 Data

This study utilized the LUMIERE dataset (n=91) [15], containing multi-parametric MRI (T1w, T2w, FLAIR, post-contrast T1w (T1c)) with longitudinal acquisitions. Automated segmentations from DeepBraTumIA (https://www.nitrc.org/projects/deepbratumia) and HD-GLIO-AUTO (https://github.com/CCI-Bonn/HD-GLIO-AUTO) models, RANO expert ratings, PyRadiomics-derived features [16] , and clinical metadata were available as part of this dataset. Due to observed segmentation inconsistencies in the LUMIERE dataset, we developed a robust deep learning (DL) based segmentation model using the BraTS 2025 Glioma Challenge dataset (comprising 1,251 pre-operative and 1,350 post-operative/post-treatment cases), which provided standardized multi-institutional MRI data with expert-validated ground truth masks for all tumor subcomponents. The proposed DL model was used for segmentation of tumor masks for the LUMIERE dataset. Details of this model are provided in the sub-section 2.2.

Since RANO criteria integrates both qualitative T2w/FLAIR data (for detecting new lesions) and quantitative T1c measurements (for bidimensional/volumetric analysis), our study utilized both FLAIR and T1c sequences. Whole tumor (WT) masks were

generated by combining the three tumor subcomponent labels: enhancing tissue (ET), surrounding non-enhancing FLAIR hyperintensity (SNFH), and non-enhancing tumor core (NETC). We developed two separate deep learning models, one for WT segmentation using FLAIR images and the other for ET segmentation using T1c images. The BraTS data was partitioned patient-wise into training (70%) and validation (30%) sets, stratified by pre-/post-operative status. Model performance was evaluated on an independent test set of 271 post-treatment cases from the BraTS dataset, ensuring robust assessment of generalizability.

2.2 DL-based segmentation model development

The automated tumor segmentation masks provided with the LUMIERE dataset exhibited notable inaccuracies. For instance, in the case of "Patient-048," the DeepBraTumIA tool miscategorized the resection cavity as a necrotic region in the post op scans. Similarly, the HD-GLIO-AUTO tool showed poor performance for the same patient, failing to accurately segment the contrast-enhancing tumor in later follow-ups (e.g., week 49). These limitations necessitated the development of more accurate, dedicated segmentation models. Consequently, we implemented and compared 3D U-Net and 3D Swin UNETR architectures for WT and ET segmentation using the BraTS dataset. To improve ET segmentations, we adopted a cascaded approach inspired by [3], where T1c images were masked using predicted WT outputs. Table 1 summarizes the key training parameters for both U-Net and Swin UNETR models. Fig. 1a shows the segmentation model development pipeline. The best performing model determined using Dice score was used for segmenting the tumor sub-components on the LUMIERE dataset for further downstream tasks following some pre-processing steps as discussed in the next sub-section.

Table 1. Training parameters used for U-Net and Swin UNETR architectures.

Parameter Category	Parameter	U-Net	Swin UNETR
Architectural	Input Patch Size	(128, 128, 64)	(128, 128, 64)
	Input Channels	1	1
	Output Channels	1	1
	Channel Sequence	(16, 32, 64, 128, 256)	N/A
	Stride Sequence	(2, 2, 2, 2)	N/A
	Feature Size	N/A	48
	Transformer Depths	N/A	(2, 2, 2, 2)
	Attention Heads	N/A	(3, 6, 12, 24)
Data & Preprocessing	Normalization	Z-Score (zero mean, unit variance)	
	Data Augmentation	Random Flipping, Random Intensity Scaling, Random Intensity Shifting	

Training	Loss Function	Weighted Dice Cross Entropy Loss
	Optimizer	AdamW
	Initial Learning Rate	1e-4
	Learning Rate Scheduler	Cosine Annealing
	Weight Decay	1e-5
	Batch Size	8
	Number of Epochs	150
	Validation Frequency	Every 5 epochs
Implementation	Hardware	GPU: NVIDIA V100 (32GB 5120 CUDA cores) CPU: 2x Intel Xeon G-6148 (20 cores 2.4 GHz)

2.3 Radiomics feature based DL-classification model development

This analysis utilized the LUMIERE dataset containing 638 longitudinal MRI scans from 91 GBM patients. After applying quality control measures, we excluded 22 scans lacking RANO ratings, 3 duplicate ratings, 14 entries without imaging data, 91 pre-operative studies, and 14 cases missing either T1c or FLAIR sequences. This resulted in 497 qualified scan timepoints. There were also a few cases with only post-op data following the pre-op data that were removed from the analysis. We treated consecutive scans as independent observations, yielding a final cohort of 377 analyzable cases with complete imaging data and RANO assessments.

The LUMIERE dataset exhibited substantial heterogeneity in image dimensions and slice counts across scans. To ensure consistency in our analysis pipeline, we registered the skull-stripped images to the SRI24 atlas templates, aligning FLAIR sequences with the T2w template and T1c sequences with the T1w template resulting in a uniform spatial resolution of 240×240×155 as shown in Fig. 1b. Finally, median-based intensity normalization was applied to standardize signal variations across the dataset.

The registered FLAIR and T1c images were processed through segmentation models to generate the WT and ET masks. From these masks, 14 shape-based and 18 first-order radiomic features were extracted (using PyRadiomics [16]) at both baseline and follow-up time points, yielding a total of 128 features per case. These features were used to classify treatment response into one of four RANO categories: CR, PR, SD, or PD. For the classification model development, an 80-20 stratified K-fold cross validation split for training and validation was used (Fig. 1c). The capability to capture key features from complex tabular data makes TabM the best for analyzing such datasets [17]. The AdamW optimizer was used to minimize a cross-entropy loss, and the training performance was evaluated using balanced accuracy. Rigorous parameter tuning was performed using Optuna [18]. Synthetic Minority Over-sampling Technique (SMOTE) was used to address the data imbalance issue (CR: 26, PD: 241, PR: 20, SD: 90).

3 Results

3.1 Segmentation model performance

Swin UNETR outperformed U-Net for segmenting both the WT and ET masks on the BraTS 2025 Glioma Challenge dataset. Moreover, using WT masked T1c images with Swin UNETR showed improved dice results compared to when using the T1c images directly. Table 2 presents the segmentation model performance results.

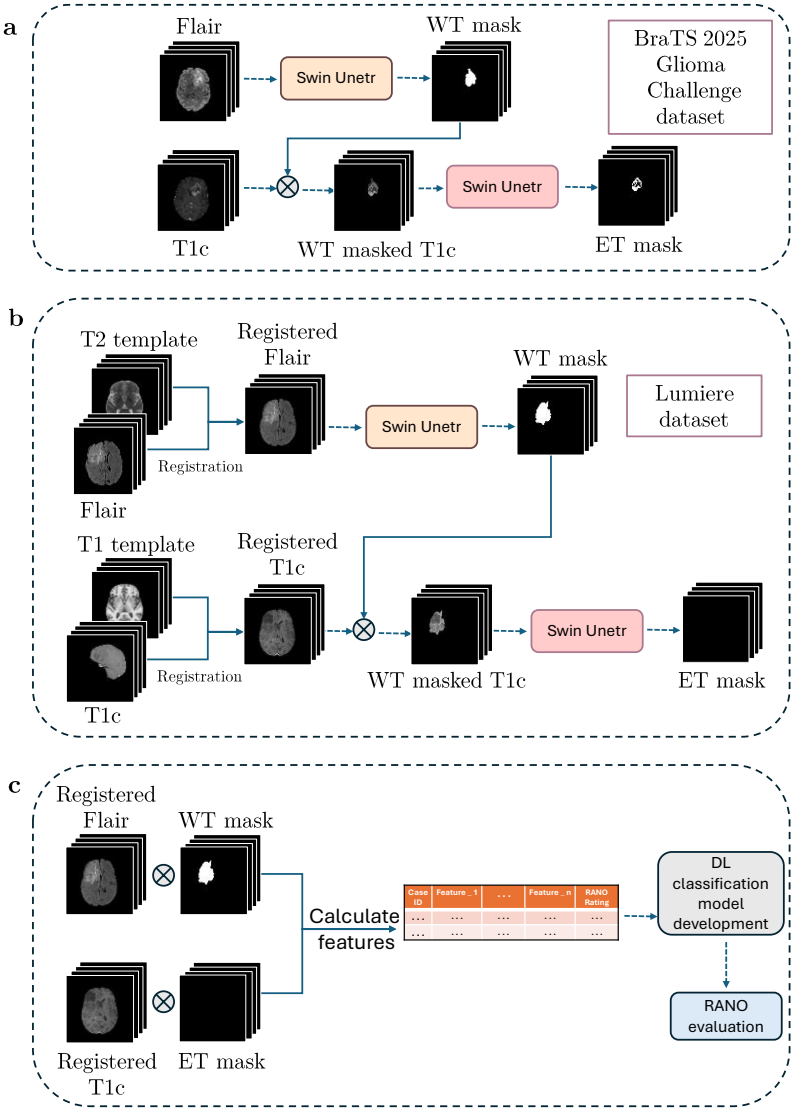

Fig. 1. The overall study design, (a) segmentation model development using the BraTS dataset, (b) Whole Tumor and Enhancing Tumor mask prediction on the LUMIERE data using the developed model from (a), and (c) radiomic feature extraction using the Lumiere data and development of the classification methods.

Table 2. Dice scores of the developed segmentation models.

Architecture	Segmentation	Metrics	Original data	Masked data
U-Net	WT	Validation Dice	0.8652	
		Test dice	0.7337 ± 0.2859	
		Median Test dice	0.8586	
	ET	Validation Dice	0.6934	0.7121
		Test dice	0.6372 ± 0.3317	0.6360 ± 0.3356
		Median test dice	0.7838	0.7948
Swin Unetr	WT	Validation Dice	**0.887**	
		Test dice	**0.7636 ± 0.2773**	
		Median test dice	**0.8811**	
	ET	Validation Dice	0.7637	**0.7944**
		Test dice	0.6832 ± 0.3317	**0.7632 ± 0.2850**
		Median test dice	0.8254	**0.8754**

Using the corresponding best performing models the WT and ET masks were predicted on the LUMIERE dataset. Fig. 2 shows the sample images along with the overlaid segmentation masks (for Patient-018), the tumor progression and the corresponding RANO ratings. Following this, the radiomic features were obtained corresponding to WT and ET masked FLAIR and T1c images, respectively. These features were then used for classification model development as discussed in the following sub section.

3.2 Classification model performance

Using the Optuna parameter tuning, the following parameters were obtained for the best performance with TabM: number of bins – 25, embedding vector dimensionality – 27, learning rate - 0.00158, L2 regularization of 0.000148, 113 ensembled sub-models, dropout of 0.3068, 191 neurons in each block, and 2 sequential blocks in the model's main backbone. Using these parameters, an average balanced accuracy of 0.6415 was obtained with 5-fold cross validation. Next, the model was trained on the whole dataset and the model weights were saved for inference. Using these saved model weights, a balanced accuracy of 0.6482 was obtained on the hidden challenge validation dataset comprising of 14 cases from 6 patients.

4 Discussion

In this study an automated pipeline for determining the RANO rating from longitudinal MRI scans was developed using a DL-based segmentation model and a radiomics based DL classifier. Swin UNETR and U-Net models were used for segmenting the tumor subcomponents from longitudinal MRI data. The results (Table 2) showed that the Swin UNETR model outperformed U-Net for both WT and ET segmentations using the BraTS 2025 Glioma Challenge dataset. Hierarchical Swin Transformer blocks used

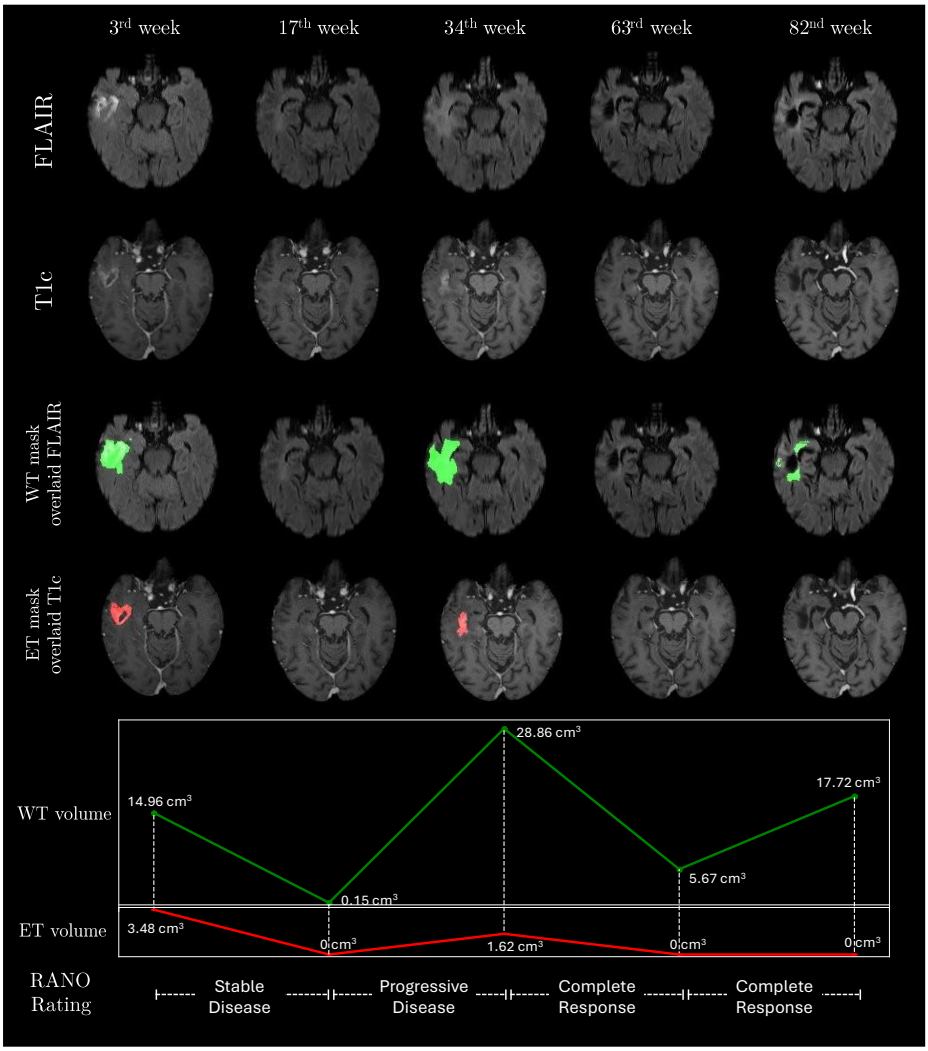

Fig. 2. Segmentation masks overlaid on the corresponding images along with the tumor subcomponent volume and the RANO ratings.

in Swin UNETR effectively capture both the global and local image context information [12]. This is advantageous over the U-Net model that is based on convolutional operations limited for taking into consideration the long-range data dependencies. Median dice scores of 0.8811 and 0.8254 were obtained for WT and ET segmentation, respectively on the original FLAIR and T1c images. Additionally, an improvement in dice score was obtained by using a cascaded method, in which the WT predictions were used to mask the T1c images to obtain the ET masks. This improved the median dice score from 0.8254 to 0.8754. Most of the other brain structures that could be falsely detected as ET were removed using this cascaded approach. Using a similar approach, the study [3] had obtained a median dice score of 0.839 on the BraTS'21 hidden test data.

The final goal of this study was to automate the RANO response assessment. This was achieved by extracting the shape-based (using WT and ET masks) and the first-order radiomic features (using FLAIR and T1c images) from the LUMIERE dataset. Features were extracted both for the baseline and the follow-up timepoints. This was aimed towards capturing quantitative changes indicative of response. Using the TabM model for classification combined with SMOTE oversampling technique, an average 5-fold cross validation balanced accuracy of 0.6415 was obtained on the LUMIERE dataset. A previous study has reported a balanced accuracy of 0.51 using the same dataset [19]. As the LUMIERE dataset is highly imbalanced with respect to its classes. In our experiments, we found that without using the SMOTE oversampling technique, the average 5-fold cross validation balanced accuracy was considerably lower at 0.5266.

The developed automated method presented in this study directly addresses the limitations of the bi-dimensional manual RANO assessments that are prone to inter as well as intra-observer variability while being time intensive. The fully automated pipeline of this study was able to output the RANO classification within 5.58 mins for a single case that includes baseline and follow-up scans using the specified hardware mentioned in Table 1. By automating both segmentation and classification, our approach offers a more reproducible and efficient alternative.

Although encouraging, the balanced accuracy that was attained also emphasizes how difficult automated RANO assessment is. Due to data restrictions in the challenge testing set, our model was unable to fully incorporate the RANO criteria, which are complex and include not only changes in tumor size but also clinical status and corticosteroid dosage. The recent RANO 2.0 update further refines these criteria, emphasizing volumetric measurements as an option and providing more specific guidance on handling non-enhancing disease and pseudoprogression. Future iterations of our pipeline should aim to incorporate these updated guidelines and additional clinical data to improve classification accuracy.

There may be major advantages to incorporating such an automated technology into the therapeutic workflow. For neuroradiologists and oncologists, it could be a decision-support tool that offers quick, numerical evaluations of therapy response. This might result in quicker and better-informed clinical judgments, which could enhance patient outcomes.

5 Limitations and Future Directions

This study has some limitations. The classification analysis was conducted on a relatively small and heterogeneous dataset (LUMIERE). While we employed robust cross-validation and a separate test set, validation on a larger, multi-institutional dataset is necessary to ensure the generalizability of our findings.

Future work should focus on several key areas. First, integrating the full spectrum of RANO 2.0 criteria, including clinical data and corticosteroid usage, is crucial for developing a more clinically relevant tool. Second, investigating the model's ability to differentiate true progression from pseudoprogression which is a major challenge in neuro-oncology, would be a significant advancement. Lastly, feature selection techniques such as MRMR (Minimum Redundancy Maximum Relevance), and LASSO (Least Absolute Shrinkage and Selection Operator) can be explored to check for model performance improvements.

6 Conclusion

In summary, this work shows that a fully automated pipeline for evaluating brain tumor response from longitudinal MRI data is both feasible and promising. Combining a Swin UNETR model for precise segmentation with a tuned DL classifier based on texture features for RANO classification, we have created a promising tool to improve the effectiveness of treatment response assessment in neuro-oncology.

7 Acknowledgements

The authors thank IIT Delhi HPC facility for computational resources.

References

1. Ostrom QT, Bauchet L, Davis FG, et al. The epidemiology of glioma in adults: a “state of the science” review. Neuro Oncol. 2014;16(7):896-913.
2. Yadav VK, Sharma S, Maurya S, et al. Presence of Fragmented Intratumoral Thrombosed Microvasculature in the Necrotic and Peri-Necrotic Regions on SWI Differentiates IDH Wild-Type Glioblastoma From IDH Mutant Grade 4 Astrocytoma. J Magn Reson Imaging.
3. Maurya S, Kumar Yadav V, Agarwal S, Singh A. Brain Tumor Segmentation in mpMRI Scans (BraTS-2021) Using Models Based on U-Net Architecture. In: International MICCAI Brainlesion Workshop. Springer; 2021:312-323.
4. Fernandes C, Costa A, Osório L, et al. Current standards of care in glioblastoma therapy. Exon Publ. Published online 2017:197-241.
5. Chinot OL, Macdonald DR, Abrey LE, Zahlmann G, Kerloëguen Y, Cloughesy TF. Response assessment criteria for glioblastoma: practical adaptation and implementation in clinical trials of antiangiogenic therapy. Curr Neurol Neurosci Rep. 2013;13(5):347.

6. Kickingereder P, Isensee F, Tursunova I, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol. 2019;20(5):728-740.
7. Chang K, Beers AL, Bai HX, et al. Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. Neuro Oncol. 2019;21(11):1412-1422.
8. Nalepa J, Kotowski K, Machura B, et al. Deep learning automates bidimensional and volumetric tumor burden measurement from MRI in pre-and post-operative glioblastoma patients. Comput Biol Med. 2023;154:106603.
9. Vollmuth P, Foltyn M, Huang RY, et al. Artificial intelligence (AI)-based decision support improves reproducibility of tumor response assessment in neuro-oncology: An international multi-reader study. Neuro Oncol. 2023;25(3):533-543.
10. Rudie JD, Calabrese E, Saluja R, et al. Longitudinal assessment of posttreatment diffuse glioma tissue volumes with three-dimensional convolutional neural networks. Radiol Artif Intell. 2022;4(5):e210243.
11. Suter Y, Schuhmacher F, Ermis E, et al. Towards Radiomics-Based Automated Disease Progression Assessment for Glioblastoma Patients. In: International MICCAI Brainlesion Workshop. Springer; 2023:36-47.
12. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. Springer; 2021:272-284.
13. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ; 2021:10012-10022.
14. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2015;9351:234-241. doi:10.1007/978-3-319-24574-4_28
15. Suter Y, Knecht U, Valenzuela W, et al. The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation. Sci data. 2022;9(1):768.
16. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77(21):e104-e107.
17. Gorishniy Y, Kotelnikov A, Babenko A. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. arXiv Prepr arXiv241024210. Published online 2024.
18. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ; 2019:2623-2631.
19. Matoso A, Passarinho C, Loureiro MP, Moreira JM, Figueiredo P, Nunes RG. Towards a deep learning approach for classifying treatment response in glioblastomas. arXiv Prepr arXiv250418268. Published online 2025.