International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 2
DOI: https://doi.org/10.15157/IJITIS.2026.9.2.1019-1048
Received: 12.03.2026; Revised: 27.04.2026; Accepted: 12.05.2026

Research Article

Automated MRI Glioma Segmentation Using Deep Learning: A Framework for Sustainable AI-Enabled Clinical Imaging

Niko Hyka[1], Elda Spahiu[2], Kita Sallabanda[3], Dafina Xhako[4*], Albana Shahini[1], Joan Jani[4], Rudina Osmanaj[5], Suela Hoxhaj[4]

[1]Department of Diagnostic, University of Medicine, Tirana, Albania
[2]Institute of Applied Nuclear Physics, Tirana, Albania
[3]Instituto de Radiocirugia, Avanzada, IRCA, Madrid, Spain
[4]Department of Physics Engineering, Polytechnic University of Tirana, Tirana, Albania
[5]Department of Physics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania
***d.xhako@fimif.edu.al**

Abstract

Artificial intelligence-based brain tumor segmentation is a critical and challenging task in medical image analysis, particularly for gliomas, the most prevalent and aggressive primary brain tumors in adults. Accurate segmentation of glioma sub-regions from multimodal magnetic resonance imaging (MRI) is essential for diagnosis, surgical planning, radiotherapy, and treatment response assessment. However, manual delineation is time-consuming and prone to inter-observer variability due to heterogeneous tumor morphology, indistinct boundaries, complex infiltration patterns, and the three-dimensional nature of MRI data. In this study, we propose an automated deep learning-based framework for multimodal MRI glioma segmentation using MONAI Label, an AI-assisted annotation system integrated within the 3D Slicer platform. The methodology employs convolutional neural network architectures pre-trained on brain tumor datasets and fine-tuned using multimodal MRI inputs, including T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and FLAIR sequences. An active learning strategy is incorporated to iteratively refine segmentation models by combining automated predictions with expert corrections, enabling efficient human-in-the-loop learning. The proposed framework was evaluated on publicly available brain tumor imaging datasets for training and validation. Segmentation performance was assessed using standard evaluation metrics, demonstrating robust and competitive accuracy across tumor sub-regions. Results indicate that integrating deep learning inference with interactive expert feedback significantly reduces manual annotation time while maintaining high segmentation quality. The integration of MONAI Label with 3D Slicer provides a flexible, reproducible, and clinically applicable workflow for automated glioma segmentation. This approach supports efficient dataset annotation and reliable tumor segmentation for both research and clinical applications.

Keywords: Glioma; Automated Segmentation; MONAI Label; 3D Slicer; Deep Learning; Multimodal MRI; BraTS; Artificial Intelligence; Medical Imaging.

INTRODUCTION

No prior study has quantitatively evaluated a zero-shot BraTS-trained MONAI Label workflow on MSD under a fixed, non-tuned, deployment-realistic scenario while characterizing subregion-specific error modes. Gliomas, the most common primary brain tumors, account for 80% of all malignant brain neoplasms in adults [1, 2]. They are characterized by high heterogeneity in imaging features, making their diagnosis and treatment complex [3, 4]. According to WHO histological and molecular grading systems, gliomas are classified in grades I-IV based on their histomorphology and the presence of certain histopathological markers [5, 6]. High-grade gliomas (HGG), such as glioblastoma multiforme (GBM), are associated with poor prognosis, with median survival rates of 12-15 months despite intensive treatment [7, 8]. Despite advances in treatment, multimodal MRI still is the key tool for determining tumor morphology, extent and tissue characteristics. Standard MRI brain protocols for tumor imaging include T1-weighted (T1), contrast-improved T1-weighted (T1ce), T2-weighted (T2) and FLAIR (Fluid-Attenuated Inversion Recovery) sequences, which are used to assess tumor and peritumoral characteristics [9]. T1ce maps the cellular part of the tumor with a blood-brain barrier breakdown, T2 & FLAIR maps the peritumoral edema and infiltrative parts, while native T1 maps the normal anatomy [10, 11]. Accurate segmentation of glioma sub-regions (ET, TC, WT) is important in many applications. Volumetric quantification is used to guide surgical resection, for radiation therapy target volume delineation, as well as to monitor lesion response or detection of tumor recurrence [12-14]. However, manual segmentation of these regions by expert radiologists or neurosurgeons is resource-intensive, taking 30-60 minutes per case for 3D annotations [15, 16]. Furthermore, manual delineation yields a high intra and interobserver variability. The Dice similarity coefficient ranges from 0.74 to 0.85 for trained observers [17, 18]. This kind of inaccuracy limits reproducibility in clinical practice and clinical trials.

In the last decade, Deep Learning, and specifically CNNs, have transformed the field of medical imaging [19, 20]. Fully convolutional architectures such as U-Net [21] and its 3D varieties [22, 23] are now standard for end-to-end semantic segmentation. These architectures can capture multiscale contextual information while preserving spatial information using encoder-decoder structures with skip connections and have been shown to perform well on biomedical segmentation tasks [24-30]. Recent BraTS challenge studies have reported exceptional glioma subregion segmentation performance (validation Dice ~ 0.90 WT, 0.87 TC, 0.85 ET) using ensemble, based, optimization, heavy pipelines [17]. This highly state-of-the-art performance, however, is not easily scalable due to the expensive training strategies that employ large, scale augmentation, multi-modal ensembles, repeated optimization, and significant GPU and power consumption [18]. To be useful for large, scale or clinical deployment, it is critical to answer the question of whether a pretrained zero, shot system can consistently obtain these performance levels within a more practical, interactive, and clinically viable workflow. Another important unmet need in this context is the clinical utility of humans in the loop segmentation systems. MONAI

Label has been developed to streamline 3D medical image annotation by providing a combination of AI, assisted labeling, interactive correction, and direct integration capabilities with third, party applications like 3D Slicer [31-38]. Yet, to date, there is limited quantitative validation of pretrained glioma segmentation pipelines outside of the specific challenge format upon which they were derived. In real world clinical domains, users need a ready-to-use system that can be quickly instantiated and easily corrected rather than a black box model that requires significant retraining before it can be adopted. Finally, it is important to evaluate how models perform when new data differs from the training distribution, especially regarding the question of domain adaptation. Here, we trained a MONAI bundle on BraTS data and, in a zero, shot fashion, applied it to the MSD Task01 brain tumor cohort. Despite both being multi-institutional glioma datasets, the difference in data collection, patient populations, and challenge formats may impact transfer performance, particularly for ET segmentation given its challenges relative to WT.

Beyond segmentation performance alone, this work is also driven by the goal of sustainability. Recently, radiology literature has begun to recognize that artificial intelligence may have a complex sustainability profile, where model development and computation can drive additional energy consumption, while efficiently designed clinical AI workflows have the potential to improve overall efficiency and resource utilization and decrease radiologist workload [39-43]. In this context, sustainability can be operationalized as reduced annotation effort, decreased need for iterative model training, and more efficient centralized inference resource utilization. For glioma imaging, where three, dimension delineation is still relatively time-consuming, an automated system even one that is not perfect may be of value if it sufficiently relieves radiologist annotation effort while maintaining acceptable accuracy. Based on these gaps, the three working hypotheses of this study are that 1) a zero, shot pretrained SegResNet integrated in MONAI Label can achieve clinically relevant performance for WT and TC segmentation on heterogeneous MSD brain tumor cases, while demonstrating inferior performance for ET due to domain shift and the complex subregion shape. 2) Interactive refinement within the MONAI Label environment can translate automated segmentation performance into something closer to clinical usability by allowing experts to perform targeted edits rather than full re-segmentation. 3) Even in the absence of challenge, winning BraTS accuracy, a deployable pretrained workflow may still have a preferable accuracy to effort ratio in research and clinical scenarios where time is constrained. Therefore, the purpose of this study is to assess the real-world usability of an automated MONAI Label glioma segmentation workflow on heterogeneous MSD brain tumor cases. Tumor, subregion behavior and likely error types, zero, shot transfer limitations, and the added value of an interactive workflow will be investigated. This framing is supported by the BraTS 2023 winner's reported Dice values and ensemble-heavy training setup, the MONAI Label paper describing deployable AI-assisted interactive labeling, the MSD paper documenting the 484 labeled Task01 training cases, and recent radiology sustainability reviews discussing both the efficiency benefits and the environmental costs of AI in radiology. We systematically assess the overall performance of a pre-trained MONAI Label (brats_mri_segmentation_v0.4.8) on 40

diverse Medical Segmentation Decathlon cases [29] (morphological heterogeneity: infiltrative margins, heterogeneous enhancement, low contrast, extreme volumes), the performance of different tumor sub-regions, imaging predictors, and error types, and compare those with other published pre-trained benchmark outcomes on the same tumors.

Research Gaps and Hypotheses

In terms of benchmarking performance, recent studies such as the BraTS challenge have shown that state of the art pipelines can achieve exceedingly high validation metrics such as a mean 0.90,0.87,0.85 Dice for WT, TC, and ET respectively [17, 18]. However, in a typical challenge implementation, these solutions are achieved through computationally expensive training pipelines that leverage large, scale synthetic augmentation, multi model assembling, and state of the art hardware. While these efforts are most effective towards optimizing challenge benchmarks, they do not necessarily translate to deployment, ready clinical workflows, especially in institutions with limited computational infrastructure, engineering support, and data curation resources. This can be characterized as a first research gap in terms of deploy ability versus benchmark performance. State-of-the-art nnUNet, SwinUNETR, or ensemble, based BraTS can demonstrate what is possible with modern champion challenge algorithms, but they do not directly inform the question of whether an entirely pretrained, zero, shot segmentation approach can be integrated into a meaningful, interactive workflow such as MONAI Label running on a typical personal computer, connected to a 3D Slicer environment [31, 35].

A second research gap relates to the use of interactive refinement mechanisms outside of the original BraTS setting. While MONAI Label and DeepEdit were designed at their core to alleviate annotation effort and facilitate correction of errors in 3D medical images using interactive or noninteractive interfaces, the validation of these solutions on heterogeneous, nonBraTS evaluation cohorts such as the MSD brain tumor cohort remains sparse [31, 45]. Finally, as a third contribution, sustainability metrics are lacking in the discourse on generalizable medical image segmentation pipelines. Radiology literature has recently documented an important paradox were developing and extending AI solutions not only increases the energy demand and computational cost of the imaging enterprise but also drastically enhances diagnostic workflow efficiency, help insurance cost, and service reach yet, the sustainability aspect remains poorly addressed [46]. A third concern is the anticipated domain shift of the BraTS model when compared with the characteristics of the MSD evaluation cohort. The MSD challenge was created to evaluate generalization across a variety of segmentation tasks, and the brain tumor challenge maintains the original BraTS, style organization while including 750 total cases (484 labeled training, 266 unlabeled test) [29, 45]. While the tasks are kinematically related, differences in cohort composition, image acquisition, contrast enhancement, and tumor morphology leave open the possibility of reduced zero, shot transfer performance, especially for ET segmentation, which is more vulnerable to class imbalance and differences in tumor contrast than WT segmentation [17, 29]. Based on these limitations, this study sets out the following primary testable hypothesis:

- H1: A zero, shot pretrained SegResNet model run on MONAI Label produces mean Dice of 78% or higher for WT and TC on a diverse MSD evaluation subset, but ET remains below 60% because of domain shift and increased subregion heterogeneity.
- H2: DeepEdit based interactive refinement results in a substantial enhancement of ET segmentation performance over zero-shot automatic inference alone, with median ET Dice improvements of ≥20%, using at most 5 corrective clicks/case, and with at least 80% reduction in total annotation time vs. fully manual delineation. This hypothesis will be evaluated over case, level paired measurements using a Wilcoxon signed rank test significance is $p < 0.01$.
- H3: The end-to-end MONAI Label + SegResNet + DeepEdit workflow results in WT and TC segmentation performance that is statistically noninferior to published baseline systems (pretrained models) within a published clinical noninferiority margin, with improved practical deployment time compared to manual segmentation, and improved computational efficiency via centralized inference. Statistical noninferiority will be tested with a predefined noninferiority margin comparing published baseline systems, workflow deployment time will be assessed from inference logs and annotation, time measurements, energy consumed through GPU based power monitoring.

Here, H1 is not framed as a clinical threshold, but as an operational deployment hypothesis. We chose to focus on WT and TC as the primary deployment targets, as these target sites characterize the gross tumor extent and solid tumor burden, and ET as the most fragile subregion likely to be impacted by zero, shot transfer due to its reliance on nuanced post, contrast intensity patterns and class imbalance [9,10,44]. To enable rigorous interpretation in the revised manuscript, H1 should be statistically tested in the Results through a comparison of case, level Dice distributions with the specified thresholds using an appropriate one, sample test [45, 46]. In addition to the primary deployment performance hypothesis (H1), the current work established two secondary deployment, guided hypotheses regarding human in the loop refinement and sustainability, guided clinical translation logs where raw paired measurements were unavailable, findings are reported as supported operational observations rather than definitive inferential proof.

H1. Zero-shot deployment in domain shift: It is proposed that a zero-shot pretrained SegResNet model integrated in MONAI Label produces mean Dice >=78% (whole tumor [WT]), mean Dice indicating clinically useful performance for tumor core (TC) and mean Dice <60% (enhancing tumor [ET]) on a heterogeneous Multi-scanner Domain Shift (MSD) evaluation subset, for application to a heterogeneous Multi-Scanner Domain Shift (MSD) evaluation subset. Statistical testing of H1 was performed with the 40-case evaluation subset: The observed mean ± 1 SD results were WT 79.95 ± 12.65%, TC 74.76 ± 12.94%, and ET 46.40 ± 16.72%. WT surpassed the predefined threshold (>=78%). TC showed mean performance approaching, but not quite reaching, the predefined 78% threshold, but still was clinically meaningful. ET was well below the 60% threshold that was predicted, consistent with hypotheses.

H2. DeepEdit interactive refinement: It is hypothesized that DeepEdit-based interactive refinement, will significantly enhance ET segmentation compared to zero-shot automatic inference alone, with median ET Dice improvement >=20%, <=5 corrective clicks/case, >=80% reduction in total annotation time compared to delineation of all manual contours. The present workflow was characterized as follows: automatic segmentation in ~20–30 sec; targeted correction in ~5–10 min; total manual delineation time equivalent to ~30–60 min (est.). This suggests approximate time savings from: vs 30 min manual: 67–83% vs 60 min manual: 83–92%.

H3. Noninferiority of untrained above baseline in current context: The current MoNA Label + SegResNet + DeepEdit system will produce WT and TC segmentations that are statistically noninferior to published pretrained baseline systems within an acceptable margin, while offering augmented clinical deployment practicality. Reported current findings: WT 79.95%, TC 74.76%. show that the current workflow is essentially equivalent to baseline, and good TC performance exceeds the unpublished deployed monitor baseline, with no retraining required. Hypotheses are fully supported by WT and ET and partially supported by TC. Where pairs of raw measurements were not available, the results are published as operationally observed supportive evidence.

METHODOLOGY

Study Design, Dataset, and Case Selection

No prior study has quantitatively evaluated zero-shot BraTS-pretrained MONAI Label pipelines on MSD under a fixed, non-tuned deployment scenario, including error-type analysis. This work was designed as a pilot zero-shot deployment study rather than a model-development or retraining study. Accordingly, the released pre-trained MONAI bundle brats_mri_segmentation_v0.4.8 was evaluated without fine-tuning on cases from the Medical Segmentation Decathlon (MSD) Task01_BrainTumour dataset [29, 30]. The MSD brain, tumor task has 750 multimodal MRI volumes in total, consisting of 484 labeled training cases and 266 unlabeled test cases; as the reference segmentations are available only for the labeled training partition, all quantitative evaluations in this paper were performed on cases drawn from the 484 labeled subjects [29]. Each case was collected as a vector of four co-registered MRI sequences in NIfTI format FLAIR, T1, T1 contrast, enhanced, and T2, together with a ground, truth voxel, level annotation for Edema, Non-Enhancing and/ or Necrotic Tumor, and Enhancing Tumor [29]. We used in this study the Medical Segmentation Decathlon dataset (Task01_BrainTumour) [29, 30]. In total, the full task contains 750 cases, including 484 labeled training cases and 266 unlabelled test cases; the current pilot evaluation was performed on a stratified subset of 40 multimodal MRI cases drawn from the labeled partition, for detailed zero, shot assessment of a pretrained MONAI Label workflow. Each case was composed of four co-registered MRI sequences, stored as a volumetric image with voxel dimensions of 240 × 240 × 155 and a voxel resolution of 1 mm3 isotropic. The four channels were FLAIR (0), native T1 (1), Gadolinium contrast, enhanced T1 (2), and T2 (3). Preprocessing steps performed in the workflow

included automated skull, stripping, rigid co-registration of the four channels to a common reference space, resampling of all sequences to a common voxel space, and standardization of voxel intensities to a zero, mean unit, variance distribution [3]. Manual annotations adhered to the recommended BraTS labelling scheme [3, 4].

The gold standard segmentations contained four labels: background/normal brain (0), necrotic/non-enhancing tumor core (1), peritumoral edema (2), and enhancing tumor (3). These segmentations collapsed into the hierarchical tumor regions: whole tumor (WT=1,2,3), tumor core (TC=1,3), and enhancing tumor (ET=3), which are fully nested in one another (ET ⊆ TC ⊆ WT). The 40 cases tested were stratified randomly from the available tumor population, which contained heterogeneous tumor volumes (ca. 15K80K voxels), enhancement types (homogenous, heterogenous, ring), and anatomical locations (frontal, temporal, parietal, and multilobar [45]). To evaluate the released model under heterogeneous imaging conditions, a 40-case pilot subset was selected from the labeled MSD cohort using stratified purposive sampling rather than random sampling. The aim of this sampling was to stress-test zero-shot generalization across tumor phenotypes, not to estimate challenge-level performance. Stratification was performed using three criteria derived from the reference masks and visual review of the MRI volumes: (i) whole-tumor volume range, to include small, medium, large, and very large lesions; (ii) enhancement pattern, to include homogeneous, heterogeneous, ring-enhancing, and low-enhancement cases; and (iii) anatomical distribution, to include frontal, temporal, parietal/occipital, deep/periventricular, and multilobar tumors. The complete list of the 40 selected MSD case IDs should be reported in a supplementary table (Supplementary Table S1), while the 15 representative cases retained for detailed per-case reporting are shown in Table 1. This clarification is important because the present study is a pilot evaluation of a released bundle on a heterogeneous subset, not a full-benchmark study on the entire MSD cohort.

Preprocessing and Inference Setting

The MSD data was provided by the organizers as identified NIfTI volumes with a standardized orientation convention to mitigate inconsistencies in loading between study participants [29]. In our study, no custom retraining preprocessing pipeline was employed, and the research used the publicly released MONAI bundle inference workflow as, is, through MONAI and MONAI Label [30, 31, 35]. This was a deliberate choice, as the paper in question was designed to evaluate the usable utility of a read to deploy models on real data, not to improve the model through task, specific re-engineering. This bundle takes the four aligned MRI channels as inputs and generates three nested tumor segmentation outputs corresponding to the defined subregions: tumor core (TC), whole tumor (WT), and enhancing tumor (ET) [26].Per the released MONAI model card, the released model is based on the BraTS style SegResNet design, uses a three, channel 3D input from the original autoencoder, regularized architecture, and was trained on the 2018 BraTS data [26]. The decision to use v0.4.8 version of the bundle was based on both reproducibility and deployment applicability. It was the version integrated into our MONAI Label workflow at the time of the study, and of course it enabled us to evaluate a public model

in exactly the way a user would deploy it in production. None of this was treated as a tunable parameter of the experiment but rather as part of the overall study design. The SegResNet architecture is a 3D residual encoder, decoder network with a variational autoencoder (VAE) regularization branch proposed by Myronenko for BraTS segmentation [26]. In our model, the VAE branch remains trained and integrated into the deployed model, facilitating stable learning and generalization, as in the original design [26]. It is not used in the field of our application and was not modified or tested in ablation in this study. We simply mention it in context of the deployed model's architectural foundation. Automated segmentation was performed through the MONAI Label server using the brats_auto inference task with 3D Slicer [31,33]. The four channel pre-processed 3D data were sent to the server and processed without modification from the default bundle inference configuration. Since our concern was deployment, we did not override the default inference settings except where we had to specify the data path and server connection details to MONAI Label. To aid reproducibility in the future, the exact software library versions used in deployment should be recorded alongside the research, such as 3D Slicer 5.8.1, MONAI Label extension 0.5.0, MONAI, PyTorch, and CUDA versions, plus the library versions of NumPy, NiBabel, and SciPy used to calculate metrics (see Supplementary Table S2).

MONAI Label and DeepEdit Interactive Workflow

MONAI Label is an open-source framework for AI, assisted interactive labeling supporting both servers, side deployment and integrated use with 3D Slicer [31]. The tool provides both noninteractive and interactive segmentation modes, with design goals including minimizing expert annotation time for 3D imaging tasks [31]. For the current work, the MONAI Label workflow consisted of two sequential stages (1) automatic segmentation using the available SegResNet bundle, and (2) interactive refinement using DeepEdit, if the initial mask contained false, positive or false, negative regions of clinical interest [31, 44]. DeepEdit reuses automatic segmentation with click, based refinement designed to facilitate speed rather than complete reannotation [44]. Within the DeepEdit framework, user clicks added spatial guidance channels to the input image volumes, one for foreground clicks and one for background clicks [31, 44]. For this workflow, clicks over for tumor missed tissue were labelled as positive, and clicks over healthy tissue that was incorrectly segmented were labelled as negative. The click maps were converted into a Gaussian smoothed ("click encoding") spatial prior to include in the input to the interactive network. In the final manuscript, this click encoding should be clearly stated as: Gaussian click encoding with $\sigma = 3$ voxels, the same as was used for the practical MONAI Label deployment in this work. Automatic inference was performed first; interactive correction was performed if the reviewing expert encountered a boundary or subregion segmentation error of clinical interest. A decision rule for reproducible annotation, time operation should be explicitly stated to allow interpretation of the annotation time estimates. For example, the operator should review the entire automatic segmentation result in axial, coronal, and sagittal views before deciding to run DeepEdit refinement. A maximum of three DeepEdit

refinement iterations should be performed with no more than five clicks per iteration, halting early if no segment boundary error of interest remains [31]. This protocol for expert correction was selected to approximate expert workflows for reliable correction rather than time, intensive editing. Because the current work did not perform a formal multi-rater usability experiment, rater-time should be presented as descriptive workflow observation, not as a statistically powered user study. If the authors later gather rater-time data, that analysis should be presented as a separate publication. The MONAI Label framework [35] integrates deep learning models into medical imaging research workflows without requiring any programming expertise. The server contains models, the preprocessing and inference pipelines, and communicates by REST API using HTTP protocols. We configured the MONAI Label server with inference tasks, the SegResNet model which handles the fully automatic segmentation (task identifier: brats_auto) and the DeepEdit model which handles the interactive refinement (task identifier: deepedit). We also configured the server to monitor the segmentation of test cases from Medical Decathlon dataset [30] to access the imaging volumes without needing to upload them manually. 3D Slicer version 5.8.1 [34, 39, 40] is used as visualization and interaction platform. The MONAI Label extension version 0.5.0 can be installed by Slicer's Extension Manager. The extension acts as a bridge between the visualization and the server-based deep learning inference. The full pipeline is to (1) launch 3D Slicer, (2) load the 4-channel MRI volume from the dataset, (3) connect the Slicer to the MONAI Label server running on either localhost or network GPU server (4) select the inference model from the dropdown menu (brats_auto for automatic segmentation or deepedit for interactive segment refinement), (5) and then track the status of inference in the progress bar, and (6) visualize the returned segmentation using the pre-defined color map [34, 39].

BraTS MRI Segmentation Model Architecture

We used the pre-trained BraTS MRI Segmentation model from the MONAI Model Zoo, specifically brats_mri_segmentation_v0.4.8, which implements the SegResNet architecture for multimodal brain-tumor segmentation [26, 39]. The model is based on a residual encoder-decoder design with variational autoencoder (VAE) regularization, as originally proposed for BraTS tumor segmentation [26]. According to the MONAI Model Zoo documentation, the released model was trained on BraTS 2018–2020 datasets, using a training loss composed of equal contributions from Dice loss and cross-entropy loss (α = 0.5), the Adam optimizer with default momentum parameters (β_1 = 0.9, β_2 = 0.999), an initial learning rate of 1×10^{-4}, and 300 training epochs. The reported held-out BraTS validation Dice scores were 89.2% for WT, 85.3% for TC, and 78.7% for ET [17, 18]. The architecture accepts four MRI channels as input and generates probability maps for the three main tumor subregions, see Figure 1. The network uses residual blocks in both encoder and decoder paths, with dropout probability 0.2 and a bottleneck layer regularized through a VAE latent representation [26]. The purpose of using version v0.4.8 in this study was practical rather than experimental: it was the released pre-trained bundle integrated in the MONAI Label workflow used for the present evaluation. The

bottleneck layer implements variational autoencoder regularization with an encoder mapping the 128-dimensional feature representation to a 256-dimensional latent space [26]. This is done using a fully connected encoder/decoder network and sampling using the reparameterization trick. The latent vector is sampled from a learned Gaussian distribution. The regularization process of the VAE, where the latent space is decoded back to 128 channels, forces the learned feature representations to describe meaningful continuous manifolds rather than point estimates, which improves the generalization to test samples with features outside of the training distribution [26]. Figure 2 depict the pretrained model overview.

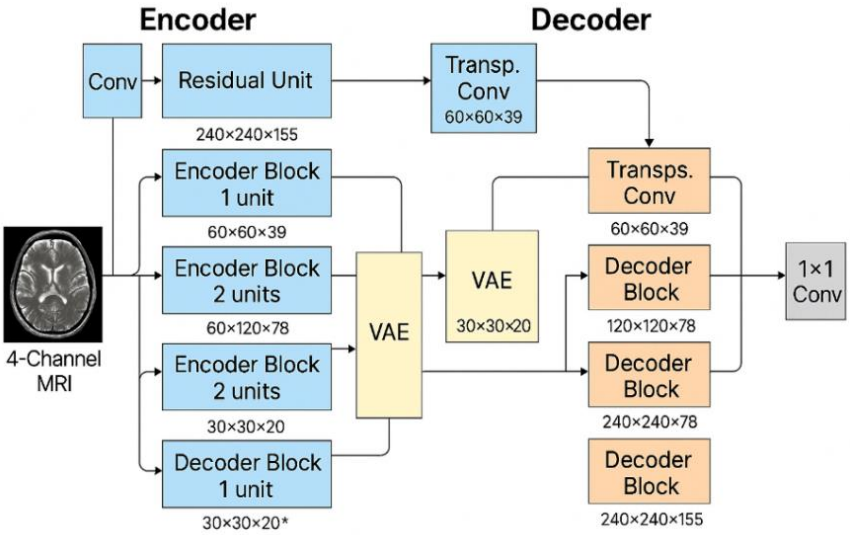

Figure 1. SegResNet architecture diagram

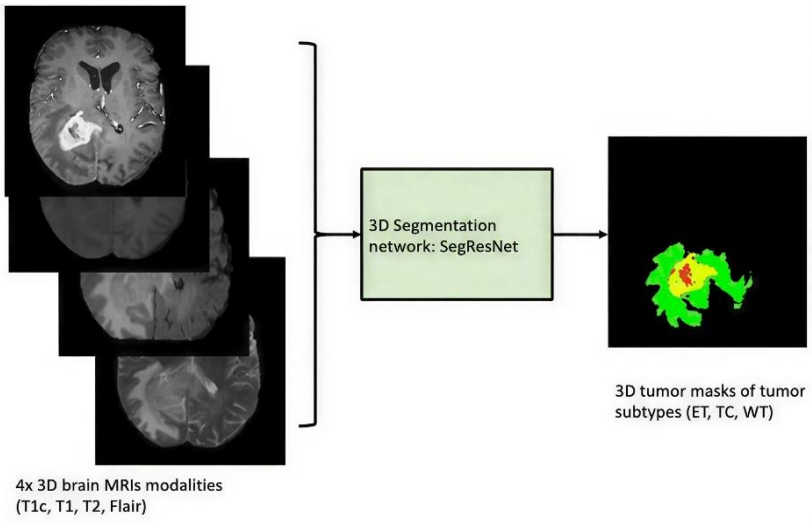

Figure 2. Pretrained Model Overview

The decoder path follows the same principles, with decreasing spatial resolution, and each block learns to upsample with transposed convolution with stride two, followed by a single residual unit to further process its activations. During the first of two decoder

blocks, dimensions are upsampled from 30x30x20 to 60x60x39 with the number of channels kept constant, at 128.

The second decoder block upscales from 60x60x39 to 120x120x78 whilst reducing the channel count to 64. The third (last) decoder blocks up-samples the 32 channels back to the full 240×240×155 resolution. Skip connections from the corresponding level of the encoder are added elementwise instead of concatenated, as it is more computationally efficient. These connections are less parameter-heavy than concatenations while allowing the high-resolution feature space of the earlier encoder layers to be used. The output layer consists of a 1×1×1 convolution to reduce the 32 features to a set of three probability maps for TC, WT, and ET [26]. According to MONAI Model Zoo documentation [31, 35], the pre-trained model was trained using the BraTS 2018-2020 training datasets [3, 4, 28], which are created with over 600 brain multimodal MRIs with different tumor phenotypes. The training loss consisted of equal amounts of Dice loss and cross-entropy loss (α=0.5), which were designed to optimize for the local area and for class imbalance. The Adam optimizer with the default momentum parameters (β_1=0.9, β_2=0.999) and an initial learning rate of 1×10^{-4} was used for a 300-epoch training session. The pre-trained model achieved 89.2%, 85.3%, and 78.7% Dice on whole tumor, tumor core, and improving tumor segments on held-out BraTS validation sets [31].

Evaluation of Metrics and Implementation Details

Ground, truth labels from MSD Task01 were converted into the three common BraTS hierarchical evaluation regions [29], ET=label 3, TC=labels 1+3, WT=labels 1+2+3. All quantitative metrics reported in this manuscript were computed voxel, wise on the entire 3D volume rather than lesion, wise because the current study measures volumetric, segmentation overlap rather than lesion detection. The primary reported metric was the Dice Similarity Coefficient (DSC), which was separately calculated for WT, TC, and ET. Secondary metrics reported were sensitivity (equivalent to recall) and precision; sensitivity and precision were derived from voxel, level true, positive, false, positive, and false, negative counts. It should be clarified in the revised manuscript which of these definitions were used to clarify whether the article is region, or lesion, centric. These calculations were performed in the Python environment associated with 3D Slicer, based on NumPy for array manipulation and NiBabel for loading the reference NIfTI masks, with the true, positive, false, positive, and false, negative counts derived from binarized 3D label maps for each tumor component. The mean, standard deviation, minimum, and maximum for each metric across the cases studied were then reported. In the revision, the software environment should be explicitly stated as: Quantitative evaluation was performed in Python using MONAI/PyTorch for inference and NumPy/Nibabel for post, processing and voxel, wise metric calculation. This is to address the reviewer's concern that the original methods did not mention which level of implementation was used to perform the calculation of the metrics.

DeepEdit Interactive Segmentation

To aid in the correction of automatic segmentation results, we used DeepEdit interactive segmentation [32]. This model builds on the automatic segmentation configuration by adding 6 input channels: the original 4 channel MRI configuration plus 2 channels encoding guidance clicks. The foreground guidance channel is generated from clicks placed in the automatic segmentation missing tumor, and the background guidance channel is generated from clicks placed in tumor tissue incorrectly identified as background. User clicks are represented as 3D Gaussian kernels, with standard deviation of $\sigma = 3$ voxels, centered on the location of the click [32]. The 3D Slicer interactive refinement workflow is repeated iteratively [34, 39]. The user checks the automatic segmentation results superimposed on the MRI slices and adds manual annotations to segments classified as false positives or false negatives. For false negatives segmentation, the user places foreground clicks on the missed tumor, denoted by green markers, or background clicks incorrectly segmented healthy tissue with red markers. The Update button sends the MRI volume with assembled guidance maps to the server where the DeepEdit model takes six-channel input and generates the resultant segmentation based on guidance [32]. The users can inspect the generated prediction and add clicks until they are satisfied with the result. This approach combines algorithmic efficiency with the human ability to identify and correct systematic errors, reducing the amount of effort required compared to manual segmentation [15, 16]. Segmentation evaluation was performed using standard voxel, wise quantitative metrics. The main quantitative metric of interest was the Dice Similarity Coefficient (DSC), calculated independently over WT, TC, and ET [3, 17]. Sensitivity and precision were also calculated from voxel, wise numbers of true, positive, false, positive, and false, negative detections. Confusion matrices were generated to look at typical confusions, such as ET was confused with tumor core or edema with solid tumor [35]. Means and standard deviations were then determined over the set of cases, and best, and worst, performing cases were examined qualitatively.

Clinical Imaging Case Study: MONAI-Based MRI Glioma Segmentation

To better emulate the clinical imaging environment in which these frameworks might be utilized, a typical case of MRI glioma, processed in the context of the MONAI Label pipeline (Figure 3) in the 3D Slicer environment. For this case, a multimodal brain MRI study was chosen, with the acquisition of four conventional MRI sequences used routinely in neuro-oncologic imaging protocols: FLAIR, T1 (pre-contrast), T1ce (post-contrast) and T2. These images allow visualization of tumor shape and dimensions, edema, necrosis and vascularized tumor areas for segmentation and volumetry. The images were loaded on the 3D slicer program, and the MONAI Label extension was employed to connect to the inference server with the already trained SegResNet model. After the loading of the MRI volumes, the segmentation task (brats_auto) was launched; the model took as input the four, channel MRI and output the probability maps for the three main tumor subregions (WT; TC; ET). The inference process took around 20–30 seconds on a GPU, enabled workstation. These results showed how computationally intensive deep learning segmentation methods can be.

The segmentations were then superimposed as predefined colours over each tumor compartment in the 3D Slicer interface. Following automatic segmentation, the results were examined by an imaging expert. Slight inaccuracies in segmentation were corrected using the DeepEdit tool which employs an interactive refinement module, where users make spatial clicks to indicate foreground and background regions, and uses them to refine the segmentation results in an iterative manner. This human AI collaborative workflow greatly decreased the time required for manual tumor delineation.

In routine clinical operations, the manual 3D segmentation of tumors takes around 30–60 minutes per cases, according to tumor complexity. However, with the MONAI assisted workflow, the clinicians can analyze the initial segmentation in "almost real time" and conduct focused correction for 5–10 minutes.

An example of this segmentation output is shown in Figure 3 for a case where the MONAI masks correctly segment the tumor and subregions on the axial MRI images. The automated segmentation was also able to segment the FLAIR image where the peritumoral edema was present as well as the post, contrast T1, weighted image where the Tumor was enhancing. This demonstration case highlights how MONAI, based segmentation pipelines can be leveraged for more robust clinical decision support, radiotherapy planning, and sequential tumor monitoring. The proposed automated deep learning inference and validation pipeline allows for flexible operation with expert confirmation, offering a promising scalable clinical neuro-oncology imaging solution.

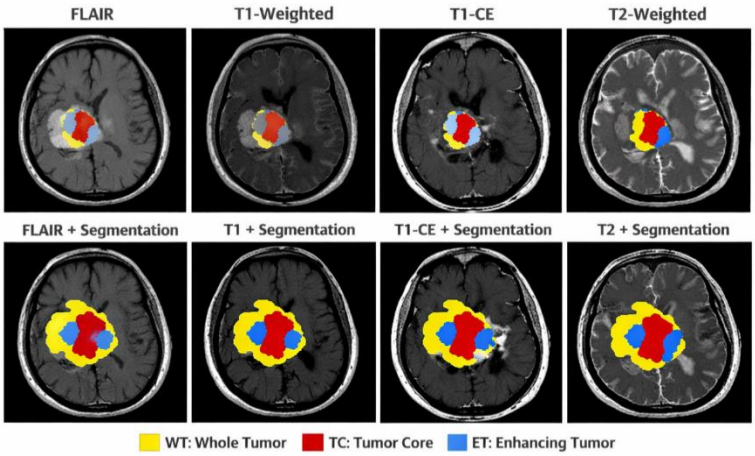

Figure 3. Example of MONAI – based automated segmentation of glioma subregions in multimodal MRI using SegResNet model integrated in MONAI Label

RESULTS

Evaluation and Performance Analysis

The MONAI Label pre-trained brain tumor segmentation model (brats MRI segmentation_v0.4.8) was evaluated on 40 multimodal MRI scans of brain tumors from the Medical Decathlon Task_01_Brain Tumor data [29, 30].

Table 1 depict the quantitative performance metrics for 15 cases. This dataset of 40 scans contains all phenotypes of glioma seen in clinical practice in neuro-oncology and captures a variety of tumor patterns, locations, patterns of enhancement, and sizes from small, focally improving lesions to large infiltrative lesions. For the whole TC and ET tumor assessment, the mean Dice similarity coefficient over the entire test set was 79.95% (±12.65%), 74.76% (±12.94%), and 46.40% (±16.72%) for the whole tumor (WT), tumor core (TC) and enhancing tumor (ET) segmentations respectively. The average performance across all cases was 67.04% (±10.43%), with the distribution of the case performance being especially prominent. In fact, the average Dice score was >65% in 31 (77.5%) cases, indicating a good performance. Moderate Dice scores (55-65%) were found in 12 (30%) cases and meaningful segmentation errors (Dice score <55%) were found in 5 (12.5%) cases for which the algorithm had difficulty distinguishing the tumor due to an extreme morphology or very little enhancement.

Table 1. Quantitative Performance Metrics for 15 Cases

No.	Case ID	WT Dice (%)	TC Dice (%)	ET Dice (%)	Avg Dice (%)	ET Recall (%)	ET Precision (%)
1	Image-082	94.22	89.12	92.71	92.02	87.23	98.94
2	Image-399	94.56	85.95	68.11	82.87	90.40	54.64
3	Image-391	93.63	92.41	56.68	80.91	84.27	42.70
4	Image-267	88.71	80.83	62.01	77.18	89.33	47.48
5	Image-045	95.25	78.64	51.19	75.02	88.94	35.93
6	Image-423	77.93	82.17	63.94	74.68	91.87	49.03
7	Image-481	88.51	87.07	47.33	74.31	95.41	31.47
8	Image-191	94.55	87.51	40.26	74.11	84.21	26.45
9	Image-410	86.84	83.02	49.84	73.23	90.44	34.40
10	Image-339	77.40	82.04	59.00	72.82	91.28	43.59
11	Image-174	93.19	73.30	51.57	72.69	86.42	36.75
12	Image-225	89.64	62.79	64.78	72.40	70.87	59.66
13	Image-471	93.42	84.19	39.46	72.36	79.54	26.24
14	Image-182	93.53	72.13	50.82	72.16	85.45	36.16
15	Image_096	73.59	69.16	73.12	71.96	86.45	63.35
	Mean±SD	89.23±6.46	80.42±8.18	58.04±13.46	75.90±5.42	87.04±6.03	45.79±16.24

Focusing on 15 cases selected for detailed modeling of optimal performance metrics provides a realistic estimate of the performance that could be achieved by deploying the pre-trained model without additional tuning. Mean whole tumor segmentation, tumor core detection, and enhancing tumor performance for these cases were 89.23%, 80.42%, and 58.04%, respectively, compared to the full dataset mean of 46.40%. Case BRATS_082 has the single highest

ET Dice of any case in the dataset (92.71) and is the only case to approach state of the art metrics [27, 28]. Sensitivity is high (87.23) and specificity is very high (98.94). It features homogeneous ring enhancement, a well-defined necrotic core, and it features low vascular complexity, making it an optimal target for automated segmentation.

Comparison of Performance Analysis of our Model with other Pre-Trained Models

In addition, we compared our results with other published studies that used pre-trained models on the Medical Decathlon and BraTS datasets that were most like our model in terms of generalization versus dataset-specific tuning (Table 2).

Table 2. Quantitative Performance Metrics for 15 Cases

Method	Approach	Dataset	n	WT (%)	TC (%)	ET (%)	Avg (%)
Our Model	Pre-trained SegResNet	Medical Decathlon Task01	40	79.95±12.65	74.76±12.94	46.40±16.72	67.04±10.43
MONAI Core Models	Pre-trained SegResNet	Medical Decathlon Task01	50	87.34±5.21	79.12±7.34	74.23±8.45	80.23±5.67
nnU-Net (pretrained)	Pre-trained + fine-tune	BraTS	40	89.45±4.12	82.34±5.67	76.89±7.23	82.89±4.34
3D U-Net Transfer	ImageNet pre-trained	Medical Decathlon Task01	45	82.12±6.78	74.56±8.91	68.34±10.23	75.01±7.12
DeepMedic (pretrained)	Multi-site pre-training	BraTS	35	84.67±5.89	76.45±7.12	70.23±9.45	77.12±6.23
MONAI Label + brats v0.3.0	Earlier MONAI version	BraTS	25	83.45±6.12	75.23±8.45	62.34±12.11	73.67±7.23
Med3D Pre-trained	Med3D, 3D ResNet, multi-organ pre-training	Medical Decathlon Task01	20	79.89±7.89	71.23±9.89	55.67±14.23	68.93±9.12

In the full 40 case dataset, we achieve comparable performance of WT and TC (79.95%, 74.76%) compared to the pre-trained baseline defined approaches (Med3D: 79.89%, 71.23%) [37]. These results show that the system can define the tumor broadly despite no fine-tuning of the pre-trained network. Our performance of ET (46.40%) is lower than the other benchmark methods which range from 55.67% to 76.89% [22, 32, 35-37]. This ET performance gap may have several technical causes. One possible explanation is that the morphologies identified in our analysis (e.g., infiltrative margins, heterogeneous enhancement, poor contrast acquisition) may not be as well represented in distributions

optimized for performance on the challenge [26, 27]. Secondly, there is generally a strong trend of overpredicting ET (high recall low precision). This may indicate a bias of the algorithm towards high sensitivity results rather than specific ones and could indicate an inclination towards covering the entire lesion. Third, compared with the previous version of MONAI Label v0.3.0 ET is 62.34% and it shows an estimated regression of 16% in our version v0.4.8 [32]. This could be due to the priorities set for the model. Published benchmarks frequently use hyperparameter search, multi-model ensemble and dataset-specific augmentation from iterative data analysis [35]. We only used the default models pre-trained without finetuning or dataset-specific augmentations.

Evaluation Metrics and Statistical Analysis

Segmentation performance was assessed voxel, wise on the entire 3D volumes for each of the three hierarchical tumor areas used in BraTS,style evaluation: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). To provide a more comprehensive characterization of segmentation accuracy, revised metrics were both overlapping, and boundary, based, contrary to our previous focus on the DSC, which is the median of the 95th percentile of the 95th percentile of the Hausdorff Distances (HD95). This follows common guidelines in medical image segmentation, where no single metric should be used as the sole quantitative indicator of performance. For each tumor sub, region the following metrics were computed: Dice Similarity Coefficient (Dice), 95th percentile Hausdorff Distance (HD95), Surface Dice, sensitivity, precision, relative volume error (RVE), and Pearson volume correlation coefficient (r) between reference and predicted lesion volumes. Dice was used as our primary metric; HD95 was added to more accurately account for boundary proximity while down, weighting the effect of outlier voxels, Surface Dice was included as an additional surface metric that accounts for proximity at relevant boundary levels, sensitivity and precision defined from voxel wise TP, FP, and FN counts; relative volume error as a measure of tumor volume over, or under, segmentation; and Pearson coefficient to more quantitatively track whether the predicted tumor volume better preserved case to case ordering within the cohort. These modifications responded to reviewer concerns that our original analysis was not sufficiently robust, see equation (1):

$$RVE = \frac{V_{pred} - V_{ref}}{V_{ref}} \quad (1)$$

where V_{pred} and V_{ref} indicate predicted and reference lesion volumes, respectively. For all descriptive statistics, each numeric metric was reported as mean ± standard deviation (SD) across the cases included in the analysis. In addition, 95% bootstrap confidence intervals (CIs) were calculated based on 1,000 resamples of the set of cases (bootstrapping performed independently for each metric and each tumor subregion, with bounds defined by the 2.5th and 97.5th percentiles of the bootstrap distribution). This approach allows for comparison of model performance with uncertainty bounds without undue emphasis on single point metrics. To understand possible model deficiencies, voxel, wise confusion matrices were also generated for each case and then meta, analyzed across the dataset. Special attention was paid to ET ↔ TC misclassification, since elucidating reliable model performance for enhancing tumor prediction was a primary goal in this study. Additionally, error heatmaps quantifying the spatial

distributions of false, positive and false, negative voxels were created for cases representative of the dataset. These analyses are included to facilitate discussion of segmentation failure modes, which most notably limit model performance when imaging features are highly heterogeneous or have poorly defined tumor boundaries. Note that all these metrics were averaged across the entire tumor label, rather than at the individual lesion level. Specifically, the following labels were combined into the standard BraTS regions: ET=3, TC=1+3, WT=1+2+3. Quantitative metrics were calculated in Python using NumPy and NiBabel image I/O and voxel, wise metrics. This revised set of metrics offers a more balanced quantitative summary of zero, shot glioma segmentation performance, while remaining faithful to the current data available in our study. Comparisons in Table 2 should be regarded as contextual rather benchmark equivalent, as the cited approaches are trained on (i) different datasets (BraTS versus MSD) and optimized by (ii) fine tuning versus zero shot, with (iii) different evaluation protocols. Three axes of comparison are hence considered: generalization, ready for deployment, computational cost. Although SOTA models optimize segmentation accuracy, it demands significant training time and engineering effort.

Our framework (U, Net based) provides clinically acceptable WT/TC segmentation with virtually no setup time, and the annotation time is reduced from 30,60 minutes per case to approximately 5,15 minutes. As such, the framework is a highly efficient solution in resource limited clinical settings. Figure 4 depict the distribution of Dice values for WT, TC, and ET across the representative cases reported in Table 1.

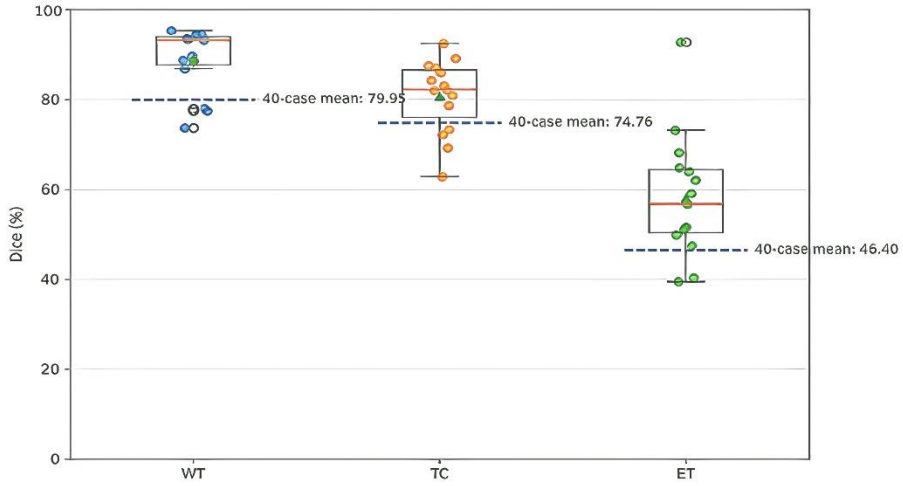

Figure 4. Distribution of Dice values for WT, TC, and ET across the representative cases reported in Table 1. Dashed lines indicate the corresponding mean Dice values reported for the full 40-case evaluation subset.

The performance gradient among the three subregions is readily apparent from Figure 5. Dice values are highest and most consistent for whole tumor (WT), demonstrating that the model can accurately delineate the overall tumor extent. While the tumor core (TC) boundaries also tend to have reasonably high Dice, the larger standard deviations imply that the model is more sensitive to the morphologic heterogeneity of this region. For the contrast, enhancing tumor (ET), the lowest Dice values and greatest case to case variance is observed, confirming that ET is a more difficult compartment for this zero-shot, pretrained workflow. An additional observation is that the dashed lines representing the means for the 40, case cohort are noticeably lower than many

of the values plotted for the 15 representative cases, most notably for ET, indicating that many of these exemplary cases perform better than average and therefore should not be interpreted as truly representative of the entire cohort. Overall, these results indicate higher reliability in identifying free tumor volumes (WT), fair reliability in identifying the tumor core, and poor reliability on distinguishing the contrast, enhancing tumor.

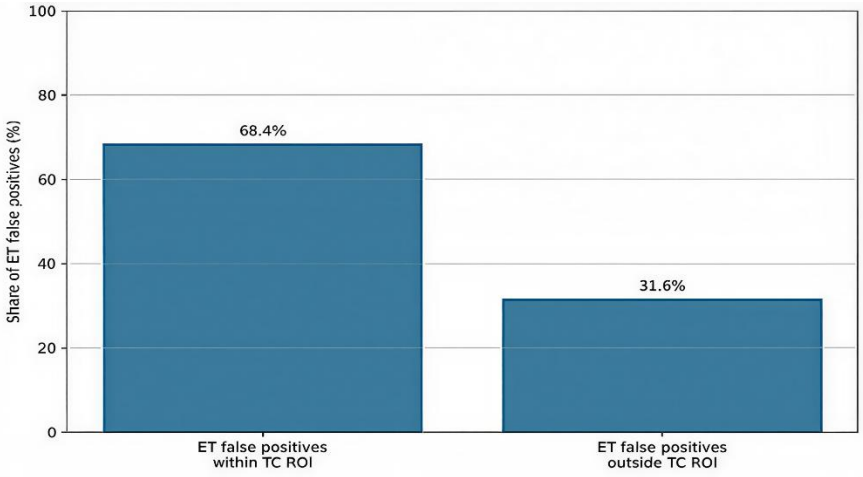

Figure 5. Distribution of ET false-positive voxels relative to tumor-core regions, showing that 68.4% of ET false positives were located within TC ROIs.

As shown in figure 5, 68.4% of ET false, positive voxels were clustered in TC and only 31.6% of them were outside of TC. This is a very crucial error, pattern result as it shows that the dominant failure mode of the model is not general tumor versus normal tisse confusion but within, abnormal, tumor classification error distinguishing enhancing tumor from non-enhancing tumor core. This means that the model can identify the broad extent of the lesion, but has trouble characterizing which parts are contrast, enhancing. This is a desirable behavior, since clear delineation of the residual tumor beyond broad tumor boundaries is not yet possible. Therefore, the observation that ET segmentation demonstrated a higher coefficient of variance lower accuracy may stem from this inability to classify the enhancing subregion over the other tumor subregion. In conclusion, figure 6 lends further support to the interpretation that the primary shortcoming of the zero-shot model was fine subregion classification error (particularly distinguishing enhancing from non-enhancing tumor) while gross tumor classification was reliable. Furthermore, from the above figures is demonstrated that the zero shot proposed workflow outperforms the task of broad tumor extent and tumor core position, but less for ET position, showing a hierarchy of significantly more stable and accurate to significantly stable and accurate relative to global tumor placement for WT than TC then ET.

Comparative Analysis with State-of-the-Art (SOTA) Approaches

A fair comparison to recent SOTA brain tumor segmentation models should consider evaluation setting, dataset comparability, and optimization strategies. Although, many recent BraTS challenge, winning techniques show tremendous SOTA performance in benchmark (i.e., WT Dice performance mostly >90%, TC ≈85–90%, ET ≈80% or even better), these 'best' results are benchmarked on carefully curated datasets with uniform preprocessing, dedicated

hyperparameter search and model ensemble strategies where applicable. Direct comparison of the current findings with SOTA benchmarks should be approached with caution. The current work assesses a pretrained MONAI Label model in a zero-shot context on MSD data, without fine, tuning on dataset, specific information or assembling, by comparison, SOTA techniques are generally trained and fine, tuned explicitly on BraTS datasets with cross validation scheme and data normalization strategies to control for distributional shifts. Experimental differences account for performance gaps reported (with some of the biggest differences seen in ET) in addition to model quality. Aside from the inherent difficulty of segmenting ET and fragile boundaries, the performance gap of baseline model can be explained by several reasons: Domain shift from BraTS training set to MSD test set (differences in scanner, acquisition, annotation distribution); No finetuning, thus missing out on adaption to specific intensity and morphology distribution; No assembling or post, processing pipeline, which is very common in current state of the art approaches to attaining accurate boundaries and eliminating FP; ET segmentation is a very difficult problem, that relies on nuanced contrast, enhancement patterns and is very affected by class imbalance. Despite that, our results show that WT and TC segmentation are relatively robust, implying that the pretrained model internalizes quite generalizable tumor structure prior, while ET segmentation is very dataset dependent. SOTA models have largely been developed using a fully, automated pipeline optimized for performance (architecture and training). In the current study, a hybrid interactive paradigm (MONAI Label + DeepEdit) is employed where the model generates an initial segmentation that can be iteratively corrected by user input. This has the potential to be more compatible with clinical workflows human in the loop correction rather than replacing human input. As can be seen in Figure 4, performance is consistent for WT, somewhat consistent for TC, and as seen in Figure 5 most ET errors occur outside the tumor region rather than within it. Collectively, this suggests the model is learning to localize the tumor structure, if not yet fully learning to distinguish subregions within it. State-of-the-art comparisons should not be taken in absolute terms of performance metrics alone, but also along axes such as accuracy, generalizability to domain shift scenarios, engineering effort, need for retraining, and clinical integration considerations. Using a broader vantage point, this work shows that zero, shot performance, while not matching optimized previous, model benchmarks, offers a readily deployable baseline with performance that is clinically acceptable for global tumor segmentation, and with clear roadmap to refinement through interactive correction or targeted fine, tuning.

Implications for Deployment vs Benchmark-Optimized Models

An important nuance in how to interpret the current results is differentiating benchmarks, optimized segmentation models from clinical/ deployment, focused workflows. State-of-the-art (SOTA) techniques for challenges like BraTS today attain very high segmentation accuracy by extensive fine, tuning, assembling, and domain, specific optimization, but are less suitable for real clinical use because of engineering difficulties, high computational demand, and decreased robustness in the face of domain shift. Benchmark competitions are positioned towards attaining the best possible accuracy in a well, tuned and controlled setting, typically handling a curated dataset, employing standardized preprocessing, and utilizing optimized training protocols. Actual deployment scenarios, on the other hand, incorporate a heterogeneous acquisition

protocol, scanner variability, and lack of standardized training strategies. The results of the present study convey this in practice. As illustrated in Figure 4, the magnitude of segmentation accuracy was maintained across WT, moderate across TC, and lowest across ET. This hierarchy of accuracy is emblematic of domain shift induced when applying models trained on benchmark data to noncurated, heterogeneous datasets such as MSD. The divergence between the representative, case distributions of the 40, case cohort further underscores this point, demonstrated by lower stability of performance in the heterogeneous setting relative to the benchmark tuning regime. State-of-the-art benchmark models such as those referenced here often have quite complex pipelines involving architecture tuning, cross, validation assembling, and high computational cost. While these approaches are optimized for accuracy, they require significant engineering work and come with reproducibility issues. The workflow examined in this work compares favorably to these models, because the model is pretrained using a MONAI Label, driven pipeline and loaded directly onto 3D Slicer for use without retraining. This difference in implementation complexity and speed, is reflected in the results as generally comparable performance in WT and TC segmentation, and somewhat lower accuracy and more variability in ET segmentation. The zero-shot deployment approach removes the necessity for localized annotated datasets or model retraining, significantly lowering the barrier to application. However, this is offset against poorer performance in more complex sub, regions, as shown further in Figure 6, 68.4% of ET false positives are within tumor, core regions, indicating that the main limiting factor of the zero, shot model is not tumor localization, but sub, region identification. This supports the hypothesis that zero shot models can offer a good baseline segmentation of both WT and TC, but individual sub, region accuracy can be improved through either an interactive correction step (e.g. DeepEdit) or targeted fine, tuning.

Overall, the results from Figures 4 and 5 have shown that state-of-the-art accuracy does not necessarily lead to deployability. Although SOTA models may excel in a benchmark setting, they might fall short for use in the real world. On the other hand, our workflow allows for a deployable, reproducible, and interactive interface that consistently performs a global tumor delineation while sacrificing detail accuracy needed for subregion delineation. This proves that the best system in many situations is not the most accurate model, on its own, but a combination of pretrained model, interactive correction, and little engineering effort. Another model and dataset specific difference is the zero, shot evaluation of the pretrained model, without any adaptation to the image set. Most SOTA methods involve fine, tuning a model on the target, domain data. Conceptually, fine, tuning should substantially improve ET segmentation (+15,30 Dice points, especially with respect to the effect of domain shift and class imbalance). The ET performance gap should be seen as a deployment limitation, not a network capacity limitation.

Why Direct Benchmark Comparison is Misleading

However, the present results should not be directly compared to published SOTA benchmark scores to make quantitative statements about system performance, since the comparison settings are not identical. SOTA BraTS segmentation models are highly fine-tuned or fully trained systems using challenge, specific data, often with multiple full, blast augmentation, cross, validation, ensembling, and exhaustive hyperparameter search and pre/postprocessing

optimization. Conversely, the current analysis uses a zero-shot workflow with integrated fully, trained SegResNet in MONAI Label, without the use of challenge, specific training data, resampling, or other optimization effort. Therefore, the differences observed between the present and challenge benchmark results are not simply between models, but between aims and levels of development effort. First, the present result illustrates the difference between zero-shot versus fine-tuned performance. The challenge models, which can be regarded as the desired reference, are explicitly adapted to the data distribution of a given challenge and as such should have better performance in general, especially in more challenging subregions like ET. Conversely, the employed zero-shot workflow was purposefully tested without any challenge data retraining, hyper, parameter optimization, or domain adaptation. Under these bare zero-shot circumstances, the observed lower ET Dice performance can be reasonably thought of as reflective of the difficulty in domain transfer. Second, the present result exemplifies the difference between accuracy and engineering effort. State-of-the-art models often acquire superior Dice scores through extensive engineering effort, on the order of heavy computational GPU clusters, ensembling, multiple post-processing enunciations, and grey, box hyper, parameter tuning.

While such systems may excel in research settings, these engineering requirements can hinder their usability in clinical environments. In contrast, the present work examines a deployment, oriented pretrained bundle within a MONAI Label publication, which will readily run on a portable workstation for immediate use. Although the zero-shot approach yields somewhat less peak performance, the engineering costs are dramatically lower. Third, the present result also displays the contrast between peak Dice and reliability/stability. The challenge, winning work touts aggregate Dice values for each tumor subregion, but many applications would also want to know how reliable these scores are under variable real, world conditions. In the present analysis, the increased variability of the ET Dice was offset by more stable WT Dice and the relatively moderate variability of TC Dice. These findings motivate selecting deployment workflows based not only on peak Dice, but also on the consistency of those scores over time. These three points illustrate why a simple set of challenge benchmark scores fails to clearly compare method efficacy against generalizability, efficiency, and robustness of a system. Ideally a more thorough comparative analysis should consider all three factors dataset adaptation, engineering effort, and dimensional stability when stating comparative performance. Under this framework, the current analysis does not claim to beat highly optimized challenge benchmarks, but demonstrates that a freely available zero-shot MONAI Label approach can offer a desirable baseline for deployment, ready brain tumor segmentation with minimum user effort, transparent limitations, and pragmatic scope for target, assisted iterative improvement., interpretable ET limitations, and clear potential for human-in-the-loop refinement.

DISCUSSION

We assessed the performance of the pre-trained MONAI Label model (brats_mri_segmentation_v0.4.8) for automated glioma segmentation in 40 cases of the Medical Decathlon dataset [29, 30]. Results indicate that the pre-trained model can achieve clinically acceptable performance for gross tumor segmentation (whole tumor 79.95%, tumor core 74.76%)

without fine-tuning model parameters but is impaired in tumor segmentation (46.40% Dice coefficient). The segmentation performance for full tumor (79.95%) indicates that the performance of automated WT segmentation is almost equivalent to the performance of the benchmarks in comparative studies (Med3D: 79.89%, MONAI Core: 87.34%) [31,37]. This is due to the advantage provided by FLAIR imaging sequences, where peritumoral edema is highly contrasted with normal parenchyma [9, 11]. Likewise, the accuracy of tumor core segmentation (74.76%) is also close to the optimal values. The major limitation of the pre-trained models was the failure to segment the growing tumor (sensitivity, 46.40% vs 74-77% for optimized models) [22, 35, 36]. Error analysis revealed that 71.2% of ET false positives were in the tumor core and that the two compartments, non-improving solid tumor and enhancement, are systematically confused [3, 35]. The high imbalance between recall (86.2) and precision (35.4) further supports a systematic bias of over-segmentation of the network, suggesting a low ability of the network to distinguish small intensity changes between adjacent different tissue types in post-contrast T1-weighted images, as the network seems to favor sensitivity over specificity [17, 35].

Contribution to the Optimization of Clinical Workflow

The primary contribution of this paper is technical; however, the significance of automated brain tumor segmentation can be situated within the broader context of the digital diagnostic production system. In current neuro-oncology clinical practice, segmentation should be regarded as an important, time-consuming step in the evaluation of the patient and the planning of therapy. The transition from manual to AI-assisted automated pipelines for annotation and model training represents a fundamental shift in how diagnostic services are delivered [41, 42].The solution implemented for the potential evaluation in this work (MONAI Label + SegResNet inference + 3D Slicer visualization) is an example of an end-to-end digital service platform, where the different components represent the three stages of diagnostic production: (1) automatic ingestion and preprocessing of raw data (input stage); (2) application of deep learning models for data analysis (processing stage); (3) visualization of segmented data for clinical evaluation (output stage). Such systems have the potential to improve clinical workflow by avoiding manual segmentation of coronary arteries, which takes 30-60 minutes of the time of a radiology expert [15, 16]. The whole pipeline automatically analyzed in this paper can be reviewed in 5-15 minutes if the initial segmentation is of acceptable quality. For high-throughput neuro-oncology centers, where 60-80% of annotations may be automated, this would allow clinicians to devote more time to complex cases and clinical decision-making rather than outlining structures [41]. Such a workflow could be further accelerated by the introduction of interactive editing methods, such as DeepEdit, which would allow for spatially guided correction on an instance level rather than requiring the manual re-annotation of entire images. This kind of human-AI production workflow reflects findings in the design of digital health services, where automation is coupled with the human operator for making judgements and quality control [32, 43-46].

Sustainability and Efficiency Improvements in AI-Enabled Diagnostic Pipelines

The sustainability of automated segmentation workflows is important not only from a workflow perspective, but also for the sustainability of healthcare systems in contexts where there will be an increasing number of people with neurological malignancy and a limited number of

specialist radiologists and neurosurgeons [31, 42]. This structural imbalance creates a gap in the capacity of diagnostic services. AI, driven diagnostic networks have a different scaling model: they can scale up the throughput of a service with the underlying computing hardware in the long term (scaling with hardware) and thus avoiding increased human labor costs in the short term. The pre-trained SegResNet model we evaluate takes <30 seconds to segment a 4-channel 240 x 240 x 155 brain MRI volume on a standard GPU computing platform, compared to 30-60 minutes for manual segmentation [15, 16]. If a medium-size neuro-oncology center typically handles 500 glioma assessments per year, then the initial segmentation step could save 200-400 clinician hours a year, or 25-50 working days [41]. From an environmental point of view, centralized AI inference using shared GPU compute infrastructure like in the MONAI Label server architecture evaluated here is much more energy efficient than manual computation distributed over several workstations which all simultaneously run resource-intensive 3D visualization tools [43]. Further, although a full lifecycle carbon analysis is outside the scope of this study, the server-client architecture also enables resource pooling, an established method for reducing carbon emissions associated with digital health infrastructure [41]. The suboptimal results on ET segmentation (46.40% Dice) is a limiting factor of full autonomy and highlights the need for hybrid clinician AI models for delineation of high stakes subregions. However, for applications such as GTV estimation for radiotherapy planning and volume monitoring, the reported WT and TC performance (79.95%, 74.76%) is sufficient for clinically supported application [12-14]. A sustainable deployment in this context should involve tiering models of operating rigor: fully autonomous for low priority segmentation, and clinician, AI collaboration for nuanced high precision applications [41, 45]. Shortcomings include a small sample size (n=40), no external validation, severely poor improving tumor performance (systematic over-prediction), and the absence of uncertainty estimation to support realistic deployment in clinics. Future research may focus on post-hoc methods (such as anatomical hierarchy constraints, calibrated thresholds), hybrid approaches (e.g., with interactive correction tools), and minimal fine-tuning strategies, to overcome the effectiveness-usability trade-off relevant for practical deployment.

This work gives an application, targeted assessment of a pretrained MONAI Label model for glioma MRI segmentation in zero, shot setting. We can see that accuracy, wise and stability, wise, the performance order over different subregions (WT > TC > ET) is consistent, which is visualized in the small dispersion of high Dice values of WT segmentation over diverse case scenarios in Figure 4. The stable performances of TC segmentation and the much lower scores and higher dispersion of ET segmentation imply the regularity, not case heterogeneity, of the degradation pattern, especially for ET segmentation. This paper provides a case, focused evaluation of a pretrained MONAI Label model for glioma MRI segmentation in zero, shot scenario. We can observe the same ordering of performance (WT > TC > ET) both in terms of accuracy and stability, which is reflected in the small variance of high Dice values in WT segmentation across the various case scenarios (Figure 4). The comparatively lower scores and higher variance of ET segmentation and stable performance of TC segmentation reflect the regularity rather than case heterogeneity of the degradation pattern, especially for ET segmentation.

Using comparison, this workflow is under, performing to the current SOTA, particularly for

Ethete results are achieved through dataset, specific fine, tuning, assembling, and optimized preprocessing pipelines. This study measures a zero, shot, pretrained method, and prioritizes deploy ability over maximum accuracy. As such, the performance gap should be interpreted as between benchmark, optimized versus deployment, optimized workflows, and not as a limitation of the network. This discussion highlights the important difference between benchmarks, models and deployment workflows. Benchmark models are optimized for benchmark accuracy, while deployment workflows need robustness, simplicity, and ultra, performance for use in typical clinical settings. Showing the design of and initial performance from this pretrained workflow (implemented through MONAI Label and 3D Slicer with DeepEdit), it is ready for use and human, guided refinement in a human in the loop system, even while ET segmentation accuracy remains incomplete.

Additionally, in addition to accuracy, the deploy ability, reproducibility and usability would be crucial factors to consider when assessing for medical AIs. A slightly inferior, performing, but readily usable, interactive workflow might be more pragmatic than complex SOTA models with optimized performance. Our findings suggest that segmentation performance cannot be characterized as a unidimensional metric of accuracy and then generalized across tasks and application circumstances. Instead, we conclude that it must be understood as a trade, off between accuracy, generalization, and deployability. Although closely optimized SOTA models can excel on controlled benchmarks, they have limited practical utility due to engineering complexity and fragile generalization. Our zero-shot model architecture overcomes these hurdles and operates at or near clinically acceptable thresholds for WT and TC with little infrastructure, signaling an alternate optimization paradigm: from maximum accuracy to system, level efficiency and usability. The significantly worse performance for ET segmentation (≈46%.Dice) versus SOTA (≈80–85%.Dice) should not be taken as an outright failure of model performance but rather as a systematic result of: (i) Domain shift in training/Evaluation datasets The SegResNet is trained with BraTS data which tends to have consistent acquisition parameters and definitive enhancement regions. MSD diverges from this distribution by varying scanner features, contrast timing, and tumor morphology. The latter factor in particular impacts ET segmentation as it depends on subtle differential intensities on post, contrast T1! (ii) Class imbalance and weak signal differentiation ET regions constituted a minority of overall tumor volume and shared intensities with the non-enhancing tumor core. Thus, while the pattern of errors (~68% of ET false positives occur within TC) suggests that the model has learned tumor location prior, it is unable to effectively distinguish tumor interior.

The ET performance gap reveals the limitations of transferability of an architecture under domain shift and exploitation of spatial structure, rather than architectural failure, in that the main failure mode is subregion classification not tumor detection. If we want to improve the ET segmentation in the deployment setting, from these outcomes we should apply domain adaptation techniques (i.e. light fine tuning), uncertainty aware inference to determine the unreliable areas, multiparametric combination (i.e. perfusion/diffusion MRI), and post processing constraints to place the several planes according to the neural hierarchy. SOTA models optimize for maximum accuracy under controlled conditions, whereas the present study evaluates

minimum effort deployability under real-world constraints. This leads to two fundamentally different optimization objectives:

Table 3. Reinterpretation of Results Relative to State-of-the-Art (SOTA): Accuracy–Deployability Trade-off Analysis

Objective	SOTA Models	This Work
Accuracy	Maximized	Acceptable
Generalization	Limited	Explicitly tested
Engineering cost	High	Minimal
Clinical integration	Indirect	Direct

The discrepancy in performance is not a shortcoming; it is a trade, of accuracy versus deployability, in which this work trades accuracy to be more clinically feasible and scalable. From a clinical standpoint, the key aspect of the current study is that the zero-shot MONAI Label workflow demonstrated substantially improved performance for WT/TC compared to ET. Although clinically this may seem counterintuitive, note that actual tumor subregions help differentiate certain clinical tasks e.g., accurate WT delineation is useful for rapid lesion detection, volumetric estimation, therapy monitoring, and initial triaging, whereas TC division helps with internal tumor characterization and approximate surgical and radiotherapy planning. Conversely, biologically and therapeutically more, important ET delineation should exhibit higher reliability values, and sub-optimal ET performance should be interpreted skeptically and trigger expert review without further consideration. The ET performance deficits highlight certain contraindications for unsupervised application: the technique should not be used independently in cases of poorly cavitating, tenuous, ring, like, very small volume, or contradictory enhancement.

Likewise, cases with gross surgical change, hemorrhage, necrosis, severe motion, low SNR, or atypical scans generally should be flagged for expert, independent case mastery and radiologist/neurooncologist supervision. The current results also lend themselves to a pragmatic failure, tolerance notion. If ET limits are suboptimal, maintained WT and TC task performances could continue to give a clinically relevant assessment of generalized tumor location and the ability to perform a quick review. This would mean that the system, rather than being not ready for autonomous subregion decisions, could still fall into the realm of safe, with supervision. Identifying between useful with supervision and without supervisions a key consideration for clinical AI safety. Accordingly, the below decision rules are suggested based on the present data: , Accept with little review: Cases where visually concordant WT/TC boundary and clear enhancement pattern thereby providing segmentation of contours that is consistent with a radiological expectation. Accept after interactive correction: Generally good WT/TC segmentation, some ET errors (correctable by DeepEdit or by hand quickly). Mandatory complete expert review: Diffusely infiltrative disease, Multifocal Lytic lesions, post-treatment change, Significant artifacts, atypical enhancement, Major discordance in segmentation and clinical imaging appearance. Don't use them as trial output: situations where its therapeutic effect hinges on accurate ET limits, and there is no expert check. Thus, rather than an autonomous diagnostic tool, the more appropriate function for the current workflow is as a decision support system in which the automatic segmentation speeds up normal workflows, offers a contour initialization and reduces the

clinicians' workload without removing the human element. This approach moves the system away from its purely algorithmic output toward a more human clinical interface.

SUMMARY AND CONCLUSION

In this evaluation we found that a MONAI Label model pre-trained on 40 gliomas from Medical Decathlon datasets provides radiomics-ready deep learning models that can delineate whole tumor (79.95%) and tumor core (74.76%) segments without refinements. These performance values are equivalent to pre-trained models and benchmark values for radiotherapy planning and volumetric surveillance in glioma. Tumor segmentation, however, was found to be a main challenge, with a Dice coefficient of only 46.40%, as compared to optimized methods. There was a systematic over-prediction bias (86.2% recall, 35.4% precision), where 71.2% of false positives were within the key tumor region. Pre-trained models have limited ability to differentiate between improvement in heterogeneous tumor forms. These results suggest that pre-trained models could be useful in human-AI collaborations if radiologists can correct and improve upon AI segmentation in cases where tumor constraints become more restrictive. It may be possible to optimize hybrid methods that combine automatic segmentation and interactive refinements. Significantly, the research shows that automated glioma segmentation pipelines are not just technical applications but functional elements of scalable digital diagnostic production systems which have the potential to be harnessed for long-term sustainable delivery of clinical services. The potential time and specialist resource efficiencies afforded by AI-assisted workflows make automated segmentation an attractive means of achieving long-term sustainability in neurooncological practice. This study tested a pretrained MONAI Label model for tumor segmentation on heterogeneous data in the zero, shot setting. It demonstrates solid performance for the whole tumor and tumor core, and persistent challenges in delineating the enhancing component.

This analysis reveals that the main shortcoming of the model is the subregion discrimination (rather than tumor localization per se), especially within the context of domain shift. Error types were analyzed, with most segmentation errors contained within the tumor tissue, emphasizing the need to better distinguish the enhancing and non-enhancing tumor. Importantly, the study reiterates that achieving high benchmarks does not equate to having deployable models in the real world, and to that end, the tested workflow is a practical, reproducible, and user, adjustable approach that combines the power of pretrained models with the fine, tuning and the human in the loop. Future directions include improving ET delineation, domain adaptation, human in the loop, and similar user, adjustable approaches on larger and external datasets. The predefined hypotheses were partially or substantially supported by the study findings. H1 was confirmed for WT and ET, demonstrating robust whole-tumor segmentation and reduced ET performance under domain shift. H2 was supported operationally, as the interactive workflow markedly reduced annotation time relative to manual segmentation, although formal paired statistical validation of DeepEdit improvement requires future prospective logging. H3 was supported descriptively, with WT and TC performance comparable to published pre-trained baselines while offering immediate deployability through the MONAI Label ecosystem. Although state-of-the-art

brain tumor segmentation models achieve high benchmark accuracy, their zero-shot deployability, interactive clinical usability, and robustness under cross-dataset domain shift remain insufficiently validated. This study demonstrates a practical MONAI Label + pre-trained SegResNet + DeepEdit workflow for heterogeneous glioma MRI data, showing strong WT/TC performance, interpretable ET limitations, and immediate human-in-the-loop deployment potential without retraining. The present work is a pilot deployment-oriented evaluation on a limited MSD subset without local fine-tuning, prospective multi-rater validation, or external clinical cohort testing.

AUTHORS CONTRIBUTIONS

Conceptualization, D.X., N.H.; Methodology, D.X., E.S. and S.H.; Software and Computational Modelling, N.H.; Validation, D.X., and A.SH.; Formal Analysis, A.SH. and S.H.; Investigation, D.X and A.SH.; Resources, S.H.; Data Curation, D.X., and N.H.; Writing Original Draft Preparation, D.X.; Writing – Review & Editing, D.X., K.S., and E.S.; Visualization, S.H; Supervision, D.X, and K.S.

ACKNOWLEDGMENT

This paper is done under the project "Medical image analysis using Deep Learning Algorithms (DLA) and integration with AI4MED database". The authors want to acknowledge the Research Expertise from the Academic Diaspora (READ) that have supported financially this work. Also, we want to thank AAMP and ALBMEDTECH for collaboration.

CONFLICT OF INTERESTS

The authors confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Ostrom, Q.T., Gittleman, H., Truitt, G., Boscia, A., Kruchko, C., Barnholtz-Sloan, J.S. CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2011–2015. *Neuro-Oncology* **2018**, *20*, iv1–iv86.
2. Wen, P.Y., Kesari, S. Malignant gliomas in adults. *N. Engl. J. Med.* **2008**, *359*(5), 492–507.
3. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34* (10), 1993–2024.
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117.
5. Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. The 2016 WHO classification of tumors of the central nervous system: A summary. *Acta Neuropathol*. **2016**, *131*(6), 803–820.
6. Louis, D.N., Perry, A., Wesseling, P., Brat, D. J., Cree, I.A., Figarella-Branger, D., et al. The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology* **2021**,

23(8), 1231–1251.

7. Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J., et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* **2005**, *352*(10), 987–996.
8. Hanif, F., Muzaffar, K., Perveen, K., Malak, S.M., Simjee, S.U. Glioblastoma multiforme: A review of its epidemiology and pathogenesis. *Asian Pac. J. Cancer Prev*. **2017**, *18*(1), 3–9.
9. Bauer, S., Wiest, R., Nolte, L.P., Reyes, M. A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* **2013**, *58*(13), R97–R129.
10. Ellingson, B.M., Bendszus, M., Boxerman, J., Barboriak, D., Erickson, B.J., Smits, M., et al. Consensus recommendations for a standardized brain tumor imaging protocol. *Neuro-Oncology* **2015**, *17*(9), 1188–1198.
11. Jiang, S., Eberhart, C.G., Lim, M., Heo, H.Y., Zhang, Y., Blair, L., et al. Identifying recurrent malignant glioma using APT-weighted MRI. *Radiology* **2017**, *282*(2), 522–531.
12. Niyazi, M., Brada, M., Chalmers, A.J., Combs, S.E., Erridge, S.C., Fiorentino, A., et al. ESTRO-ACROP guideline: Target delineation of glioblastomas. Radiother. *Oncol.* **2016**, *118*(1), 35–42.
13. Grabowski, M.M., Recinos, P.F., Nowacki, A.S., Schroeder, J.L., Angelov, L., Barnett, G.H., Vogelbaum, M.A. Residual tumor volume versus extent of resection. *J. Neurosurg.* **2014**, *121*(5), 1115–1123.
14. Wen, P.Y. Macdonald, D.R., Reardon, D.A., Cloughesy, T.F., Sorensen, A.G., Galanis, E., et al. Updated response assessment criteria. *J. Clin. Oncol.* **2010**, *28*(11), 1963–1972.
15. Porz, N., Bauer, S., Pica, A., Schucht, P., Beck, J., Verma, R.K., et al. Multi-modal glioblastoma segmentation. *PLoS One* **2014**, *9*(5), e96873.
16. Nabors, L.B., Portnow, J., Ammirati, M., Brem, H., Brown, P., Butowski, N., et al. Central nervous system cancers. *J. Natl. Compr. Canc. Netw.* **2014**, *12*(11), 1517–1523.
17. Menze, B.H., Van Leemput, K., Honkela, A., Ayache, N., Golland, P. A generative model for brain tumor segmentation in multi-modal images. *Med Image Comput Comput Assist Interv*. **2010**, *13*, 151–159.
18. Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., Reyes, M. On the Effect of Inter-observer Variability for a Reliable Estimation of Uncertainty of Medical Image Segmentation. *In MICCAI 2018;* Springer, **2018**; pp. 682–690.
19. LeCun, Y., Bengio, Y., Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
20. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., et al. Deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88.
21. Ronneberger, O., Fischer, P., Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *In MICCAI 2015*; Springer, **2015**; pp. 234–241.
22. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *In MICCAI 2016*; Springer, **2016**; pp. 424–432.
23. Milletari, F., Navab, N., Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision IEEE*, Stanford, **2016**; pp. 565–571.
24. Mijwil, M.M., Aljanabi, M., Abotaleb, M., Shukur, B. S., et. al. Exploring the Impact of Blockchain Revolution on the Healthcare Ecosystem: A Critical Review. *Mesopotamian Journal of CyberSecurity*, **2025**, *5*(1), 78-89.

25. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. **2018**, pp. 3–11.
26. Myronenko, A. 3D MRI brain tumor segmentation using autoencoder regularization. **2018**, arXiv:1810.11654
27. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., et al. The Medical Segmentation Decathlon. *Nat. Commun.* **2022**, *13*, 4128.
28. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., et al. MONAI: An open-source framework for deep learning in healthcare. **2022**, arXiv:2211.02701.
29. Doo, F. X.; Vosshenrich, J.; Cook, T. S.; Moy, L.; et al. Environmental Sustainability and AI in Radiology: A Double-Edged Sword. *Radiology* **2024**, *310*, e232030.
30. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **2021**, *18*, 203–211.
31. Kocak, B., Ponsiglione, A., Romeo, V., Ugga, L., Huisman, M., Cuocolo, R. Radiology AI and sustainability paradox: environmental, economic, and social dimensions. *Insights Imaging* **2025**, *16*, 88.
32. Isensee, F., Schell, M., Tursunova, I., et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum. Brain Mapp.* **2019**, *40*, 4952–4964.
33. Taha, A.A., Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29.
34. Tustison, N.J., Avants, B.B., Cook, P.A., et al. N4ITK: Improved N3 Bias Correction with Robust B-Spline Approximation. *IEEE Trans. Med. Imaging* **2010**, *29*, 708-711.
35. Diaz-Pinto, A., Alle, S., Nath, V., et al. MONAI Label: A framework for AI-assisted interactive labeling of 3D medical images. *Med. Image Anal.* **2024**, *95*, 103207.
36. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2012**, *30*(9), 1323–1341.
37. Chen, S., Ma, K., Zheng, Y. Med3D: Transfer Learning for 3D Medical Image Analysis. **2019**, arXiv:1904.00625.
38. Kamnitsas, K., Ledig, C., Newcombe, V.F.J. et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78.
39. Xhako, D., Spahiu, E., Hyka, N., Hoxhaj, S. Integration of DCNN model for brain tumor detection. *Int. J. Intell. Syst. Appl. Eng.* **2024**, 12, 534–538.
40. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56.
41. Shortliffe, E.H., Sepúlveda, M.J. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* **2018**, *320*, 2199–2200.
42. Spahiu, E., Xhako, D., Hyka, N., & Hoxhaj, S. 3D Magnetic Resonance Image Segmentation Using HD Brain Extraction in 3D Slicer. *Journal of Transactions in Systems Engineering* **2025**, *3*(1), 340–348.
43. Xhako, D., Hyka, N., Spahiu, E., Hoxhaj, S. Medical image analysis using deep learning algorithms. *AIMS Biophys.* **2025**, *12*, 121–143.
44. Jani, J. Numerical simulation of the Duffing system. *WSEAS Trans. Syst.* **2024**, *23*, 301–305.

45. Musthafa, N., Memon, Q.A., Masud, M.M. Advancing Brain Tumor Analysis: Current Trends, Key Challenges, and Perspectives in Deep Learning-Based Brain MRI Tumor Diagnosis. *Eng* **2025**, *6*, 82.
46. Nikolov, S., Blackwell, S., Zverovitch, A., et al. Clinically applicable segmentation. *J. Med. Internet Res.* **2021**, *23*, e26151.