Toward Digital Twin-Ready Brain Tumor Segmentation: A Multi-Pillar Comparison of Transformer and Convolutional Models

Tanmoy Debnath
Charles Sturt University
Bathurst, New South Wales, Australia
tdebnath@csu.edu.au

Miroslaw Narbutt
Technological University Dublin
Dublin, Ireland
miroslaw.narbutt@tudublin.ie

Abstract

Patient-specific modeling pipelines in neuro-oncology increasingly require segmentation models that preserve not only spatial overlap, but also boundary fidelity and volumetric reliability. In this study, we present a rigorous comparative evaluation of three state-of-the-art brain tumor segmentation frameworks, nnU-Net v2, the Universal Model, and VSmTrans, on the MSD Task01 Brain Tumor benchmark using five-fold cross-validation, a multi-pillar metric framework, and non-parametric statistical testing. Across Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) subregions, VSmTrans delivers the strongest overall performance, achieving the best balance of overlap accuracy, boundary precision, and cohort-level volumetric agreement. However, the analysis also reveals a clinically important trade-off: although VSmTrans provides superior geometric fidelity, nnU-Net v2 exhibits lower volumetric bias for enhancing tumor estimation. These findings demonstrate that overlap-only evaluation is insufficient for downstream patient-specific modeling, where geometric and volumetric errors may propagate into longitudinal monitoring and digital twin-ready analytical workflows. This study provides an evidence-based benchmark for selecting segmentation architectures that are better suited to patient-specific and digital twin-ready brain tumor analysis.

CCS Concepts

• **Applied computing** → **Life and medical sciences**; *Life and medical sciences*; • **Computing methodologies** → **Neural networks**.

Keywords

Brain Tumor Segmentation, Deep Learning, Vision Transformers, Artificial Intelligence, Volumetric Analysis.

ACM Reference Format:
Tanmoy Debnath and Miroslaw Narbutt. 2026. Toward Digital Twin-Ready Brain Tumor Segmentation: A Multi-Pillar Comparison of Transformer and Convolutional Models. In *Proceedings of SIGKDD (BIOKDD, SIGKDD'26).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

BIOKDD, SIGKDD'26, Jeju, Korea
2026. ACM ISBN 978-1-4503-XXXX-X/2026/XX
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent advances have established deep learning (DL) as the dominant paradigm for brain tumor segmentation, shifting attention from basic architecture design toward clinical applicability, computational efficiency, and interpretability [15]. Although U-Net variants remain the prevailing baseline for delineating complex tumor boundaries [13], current research increasingly aims to reduce dependence on full multiparametric MRI protocols. Recent studies have shown that competitive performance can be achieved using only T1c and FLAIR sequences, thereby supporting more streamlined clinical workflows [7]. In parallel, explainable AI (XAI) methods are being incorporated into segmentation pipelines to improve transparency for radiologists [16], while lightweight adaptations of YOLO have demonstrated the feasibility of near-real-time segmentation for neurosurgical planning [18].

Within this broader landscape, several foundational DL architectures now define the state of the art in medical image segmentation. Among them, nnU-Net v2 [8] serves as the established convolutional baseline, offering a self-configuring and robust framework that minimizes manual design bias. The Universal Model [11] extends this direction through vision-language alignment, enabling text-guided anatomical and tumor label encoding that supports stronger generalization across heterogeneous datasets and imaging modalities. VSmTrans [12], by contrast, represents a hybrid Transformer design that combines convolutional locality with Swin-based attention to capture long-range dependencies while controlling computational overhead. These architectural differences are particularly relevant in brain tumor segmentation, where automated delineation of glioma subregions remains difficult because of marked anatomical heterogeneity, irregular morphology, and severe class imbalance.

Accurate segmentation of glioma subregions is fundamental for surgical planning, radiotherapy, longitudinal monitoring, and emerging patient-specific computational workflows.

Consequently, segmentation quality should be evaluated beyond overlap accuracy alone, since similar Dice scores may conceal substantial differences in boundary fidelity and volumetric reliability [10].

Motivated by these considerations, we present a systematic comparative evaluation of three state-of-the-art segmentation frameworks—nnU-Net v2, the Universal Model, and VSmTrans—on the MSD Task01 Brain Tumor benchmark [2]. Our study employs a standardized five-fold cross-validation protocol, a multi-pillar evaluation framework spanning overlap, boundary, and volumetric metrics, and non-parametric statistical testing to characterize performance differences across Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) subregions. The principal contributions of this manuscript are as follows:

- We present a systematic comparative benchmark of three state-of-the-art brain tumor segmentation frameworks on MSD Task01 Brain Tumor using a standardized five-fold cross-validation protocol.
- We introduce a multi-pillar evaluation framework that jointly evaluates overlap, boundary, volumetric, and classification metrics.
- We show that VSmTrans achieves the strongest overall subregion performance across Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET), while also identifying a clinically important trade-off between geometric precision and volumetric reliability, particularly for ET segmentation.
- We provide a reproducible benchmarking pipeline with standardized preprocessing, archived checkpoints, and non-parametric statistical testing.

2 Related Work

2.1 Convolutional Architectures

Recent reviews emphasize that accuracy, robustness, interpretability, and computational feasibility are all relevant when models are considered for medical imaging workflows [1, 15]. The U-Net [17] established the encoder-decoder paradigm with symmetric skip connections as the dominant framework for volumetric medical image segmentation, and its variants remain the primary baseline across glioma benchmarks. nnU-Net [8] extended this design by automating dataset-specific configuration, encompassing preprocessing heuristics, patch-size selection, network topology, and post-processing, thereby producing a self-configuring baseline that consistently matches or surpasses manually tuned solutions without architecture modification. Subsequent variants have introduced residual connections [13], attention gates, and multi-scale feature integration to improve boundary mapping in morphologically heterogeneous tumor subregions. Recent work has additionally explored reducing reliance on full multiparametric MRI protocols: competitive segmentation performance has been demonstrated using only T1c and FLAIR sequences [7]. Lightweight YOLO-adapted architectures have further demonstrated real-time viability for surgical navigation [18]. Explainability-aware architectures incorporating LIME or Grad-CAM techniques [16] represent a parallel thread prioritizing clinical transparency alongside accuracy.

2.2 Transformer and Hybrid Architectures

Vision Transformers introduced global self-attention mechanisms that overcome the limited receptive fields of convolutional operators [5]. UNETR [6] demonstrated that pure transformer encoders combined with CNN decoders can achieve competitive volumetric segmentation, though at substantially higher computational cost than equivalent convolutional baselines. The Universal Model [11] extends prior segmentation frameworks by incorporating Contrastive Language-Image Pre-training (CLIP) to enable text-guided encoding of anatomical and tumor labels. This strategy improves semantic alignment across heterogeneous datasets. Hybrid architectures coupling convolutional inductive biases with global attention have emerged as a pragmatic compromise. VSmTrans [12] implements this hybrid design strategy by integrating variable-shape window multi-head self-attention and lightweight parallel convolutional branches. By coupling global dependency modeling with local feature extraction, it achieves competitive 3D segmentation performance with lower overhead than pure transformer models. Systematic reviews confirm that hybrid models consistently outperform single-architecture approaches in brain tumor classification [9, 15], supporting the multi-architecture evaluation strategy taken here.

2.3 Benchmarking and Evaluation

The Medical Segmentation Decathlon [2] established MSD Task01 as a standardized glioma benchmark enabling reproducible cross-study comparison across architectures and modalities. Evaluation methodology has received increasing attention as recent reviews emphasize that overlap-only measures such as the Dice Similarity Coefficient (Dice) provide an incomplete assessment of segmentation quality. Boundary accuracy, volumetric agreement, robustness, and clinically meaningful failure modes should also be considered when comparing segmentation models [1, 10, 14]. Recent meta-analytic evidence further indicates that segmentation performance differs substantially by tumor subregion and is associated with the use of 3D architectures and multiparametric MRI inputs [4]. The multi-pillar evaluation framework employed here, combining overlap, surface-distance,

volumetric, and statistical analyses, directly addresses these recommendations.

3 Methodology

3.1 Dataset

MSD Task01 Brain Tumor contains 750 multimodal MRI volumes: 484 labeled training scans and 266 challenge test scans whose labels are withheld [2]. Each scan provides four co-registered modalities: native T1, post-contrast T1c, T2, and FLAIR, acquired at $240 \times 240 \times 155$ voxels before challenge preprocessing. Ground-truth labels define three nested tumor subregions: Whole tumor (WT, labels 1, 2, and 4), tumor Core (TC, labels 1 and 4), and Enhancing tumor (ET, label 4 only), where labels correspond to necrotic and non-enhancing core (1), peritumoral edema (2), and gadolinium-enhancing tumor (4). Five-fold cross-validation was applied exclusively to the 484 training scans ($\approx$ 387 training and $\approx$ 97 validation subjects per fold), as summarized in Table 1. The held-out test set was not used for model development.

Table 1: MSD Task01 dataset summary and five-fold cross-validation split sizes.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	387	387	388	388	388
Validation	97	97	96	96	96
Modalities	T1, T1c, T2, FLAIR				
Subregions	WT, TC, ET				

3.2 Preprocessing, and Model Training

Preprocessing was standardized across models while respecting framework-specific biases. nnU-Net v2 and VSmTrans pipelines applied RAS reorientation, resampling, intensity clipping (0.5–99.5%), z-score normalization, and foreground-aware cropping. The Universal Model used channel-wise z-score normalization and fixed ROI extraction (128^3 voxels). Online data augmentation (random flips, rotations, scaling, intensity perturbations) was applied consistently to enhance generalization. All models were implemented in PyTorch and trained on an NVIDIA RTX 5090 GPU (32 GB VRAM) with Automatic Mixed Precision. Training lasted up to 1000 epochs with early stopping (patience = 100, metric: validation Dice). Optimization followed framework recommendations: nnU-Net v2 used SGD with Nesterov momentum (0.99) and polynomial LR decay (1×10^{-2}), VSmTrans used Adam (1×10^{-4}), and the Universal Model used AdamW (2×10^{-4}). Experiments ran on Ubuntu 22.04 LTS, Python 3.11.5, PyTorch 2.1.0, and cuDNN 9.5, with all checkpoints archived for reproducibility (SHA–256 hashes).

3.3 Evaluation Metrics

To provide a comprehensive evaluation of model performance across WT, TC, and ET tumor subregions, the fifteen computed metrics were organized into four clinically distinct evaluation pillars [10]:

- **Overlap and Voxel-Wise Metrics:** Dice Similarity Coefficient (Dice), Intersection over Union (IoU), Precision, Recall (Sensitivity), Matthews Correlation Coefficient (MCC), Cohen's Kappa Coefficient (Kappa), and Area Under the Receiver Operating Characteristic Curve (ROC_AUC) were used to quantify pixel-level correspondence, regional alignment, and classification robustness. Matthews Correlation Coefficient and Cohen's Kappa were specifically included to mitigate the effects of class imbalance in small tumor sub-regions and to better capture detection fidelity.
- **Surface Distance (Geometric) Metrics:** The 95^{th} percentile Hausdorff Distance (HD95) and Average Symmetric Surface Distance (ASSD) were employed to assess boundary localization accuracy and geometric fidelity. Unlike overlap-based measures, these metrics explicitly penalize boundary outliers and shape inconsistencies, which are critical in clinical contexts such as surgical intervention and radiotherapy planning.
- **Volumetric Estimation and Agreement Metrics:** Volume Similarity (VS), Relative Volume Error (RVE), Volumetric Similarity Error (VSE), and Pearson Correlation Coefficient (Pearson) were used to evaluate longitudinal tumor burden tracking, with emphasis on accurate absolute volume quantification rather than solely spatial co-location. Negative RVE values indicate under-segmentation, while positive values reflect over-segmentation. This pillar combines volume-based accuracy and correlation analysis to assess clinical volumetric reliability.
- **Supplementary Classification Metrics:** Overall Accuracy (Accuracy) and Negative Predictive Value (NPV) were included for completeness, providing a global voxel-wise performance overview. They are not interpreted because of severe background imbalance.

For overall model comparison, an aggregated summary was computed by averaging each evaluation metric across the three tumor regions (WT, TC, and ET). The associated SD reflects the variability among the three regional mean values.

3.4 Statistical Methods

Statistical analyses used paired per-subject evaluation metrics from validation subjects in each fold of the five-fold cross-validation protocol. For each segmentation metric and tumor subregion (WT, TC, and ET), observations from all

Table 2: Region-wise Quantitative Comparison on MSD Task 01 (Brain Tumor). Evaluation metrics are grouped by domain logic. Arrows represent the preferred direction of each metric and best results are highlighted in bold.

Metrics	WT (Whole Tumor)			TC (Tumor Core)			ET (Enhancing Tumor)		
	nnUNet	Universal	VSmTrans	nnUNet	Universal	VSmTrans	nnUNet	Universal	VSmTrans
1. Overlap and Voxel-Wise Metrics (Higher is better)									
Dice(↑)	0.802 ± 0.016	0.759 ± 0.002	**0.807 ± 0.013**	0.625 ± 0.018	0.555 ± 0.022	**0.626 ± 0.021**	0.785 ± 0.023	0.727 ± 0.010	**0.786 ± 0.022**
IoU(↑)	0.683 ± 0.018	0.627 ± 0.002	**0.690 ± 0.015**	**0.491 ± 0.017**	0.418 ± 0.016	0.489 ± 0.018	0.687 ± 0.027	0.625 ± 0.010	**0.687 ± 0.025**
Precision(↑)	**0.835 ± 0.018**	0.756 ± 0.011	0.808 ± 0.011	0.685 ± 0.026	**0.692 ± 0.021**	0.685 ± 0.031	0.797 ± 0.016	0.755 ± 0.033	**0.816 ± 0.011**
Recall(↑)	0.791 ± 0.030	0.792 ± 0.012	**0.827 ± 0.028**	**0.649 ± 0.039**	0.531 ± 0.042	**0.649 ± 0.065**	0.804 ± 0.033	0.779 ± 0.017	**0.817 ± 0.043**
MCC(↑)	0.806 ± 0.015	0.765 ± 0.002	**0.811 ± 0.012**	0.643 ± 0.017	0.583 ± 0.016	**0.645 ± 0.019**	0.791 ± 0.022	0.764 ± 0.014	**0.815 ± 0.027**
Kappa(↑)	0.801 ± 0.016	0.758 ± 0.002	**0.806 ± 0.013**	0.624 ± 0.018	0.554 ± 0.022	**0.625 ± 0.021**	0.785 ± 0.023	0.727 ± 0.010	**0.786 ± 0.022**
ROC_AUC(↑)	0.524 ± 0.005	0.896 ± 0.006	**0.913 ± 0.014**	0.522 ± 0.003	0.765 ± 0.021	**0.823 ± 0.032**	0.524 ± 0.001	0.887 ± 0.008	**0.909 ± 0.020**
2. Surface Distance Metrics (Lower is better)									
HD95(↓)(mm)	6.379 ± 0.609	10.168 ± 1.742	**6.248 ± 0.698**	7.745 ± 1.271	9.004 ± 0.886	**7.266 ± 0.852**	4.252 ± 1.501	9.071 ± 2.061	**4.024 ± 0.940**
ASSD(↓)(mm)	**1.397 ± 0.059**	2.152 ± 0.266	1.408 ± 0.065	1.735 ± 0.139	2.173 ± 0.188	**1.672 ± 0.208**	1.089 ± 0.243	2.586 ± 0.670	**1.047 ± 0.187**
3. Volumetric Estimation & Agreement Metrics									
VS(↑)	0.911 ± 0.016	0.887 ± 0.003	**0.915 ± 0.009**	0.782 ± 0.037	0.740 ± 0.029	**0.793 ± 0.020**	**0.874 ± 0.010**	0.828 ± 0.013	0.868 ± 0.018
RVE(↓)	**-0.026 ± 0.052**	0.122 ± 0.035	0.066 ± 0.046	0.765 ± 0.772	**-0.096 ± 0.150**	0.477 ± 0.343	**0.086 ± 0.118**	0.801 ± 0.327	0.149 ± 0.268
VSE(↓)	**0.174 ± 0.024**	0.266 ± 0.008	0.190 ± 0.018	1.119 ± 0.750	**0.470 ± 0.049**	0.814 ± 0.271	**0.256 ± 0.107**	1.011 ± 0.293	0.345 ± 0.249
Pearson(↑)	0.808 ± 0.012	0.765 ± 0.002	**0.811 ± 0.012**	0.644 ± 0.016	0.583 ± 0.016	**0.645 ± 0.019**	**0.817 ± 0.025**	0.764 ± 0.014	0.815 ± 0.027
4. Supplementary Classification Metrics									
NPV(↑)	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00
Accuracy(↑)	0.998 ± 0.00	0.997 ± 0.00	0.998 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00	0.999 ± 0.00

Table 3: Overall performance summary across the three tumor regions. Each value is computed from the corresponding WT, TC, and ET results in Table 2; the reported standard deviation (SD) represents the variability among the three tumour-region means. Best results are highlighted in bold.

Metric	nnUNet	Universal	VSmTrans
Dice(↑)	0.737 ± 0.098	0.680 ± 0.110	**0.740 ± 0.099**
IoU(↑)	0.620 ± 0.112	0.557 ± 0.120	**0.622 ± 0.115**
Precision(↑)	**0.772 ± 0.078**	0.734 ± 0.036	0.770 ± 0.073
Recall(↑)	0.748 ± 0.086	0.701 ± 0.147	**0.764 ± 0.100**
MCC(↑)	0.747 ± 0.090	0.704 ± 0.105	**0.757 ± 0.097**
Kappa(↑)	0.737 ± 0.098	0.680 ± 0.110	**0.739 ± 0.099**
ROC_AUC(↑)	0.523 ± 0.001	0.850 ± 0.073	**0.882 ± 0.050**
HD95(↓)	6.125 ± 1.760	9.414 ± 0.654	**5.846 ± 1.658**
ASSD(↓)	1.407 ± 0.323	2.303 ± 0.245	**1.376 ± 0.314**
VS(↑)	0.856 ± 0.066	0.818 ± 0.074	**0.859 ± 0.062**
RVE(↓)	0.275 ± 0.428	0.275 ± 0.468	**0.231 ± 0.217**
VSE(↓)	0.516 ± 0.524	0.582 ± 0.385	**0.450 ± 0.325**
Pearson(↑)	0.756 ± 0.097	0.704 ± 0.105	**0.757 ± 0.097**
NPV(↑)	0.999 ± 0.000	0.999 ± 0.001	0.999 ± 0.000
Accuracy(↑)	0.998 ± 0.001	0.998 ± 0.001	0.999 ± 0.001

validation subjects were pooled across folds while preserving the correspondence between predictions from different models for the same subject. For descriptive reporting, the

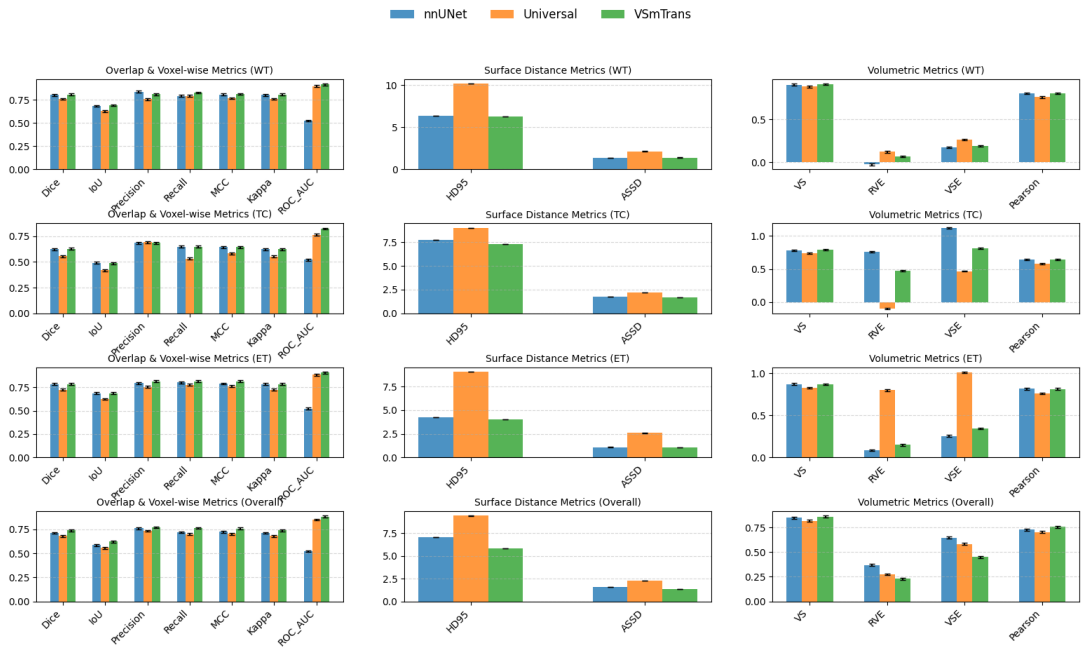

Figure 1: Performance of the 3 models across subregions. Metrics are grouped into *Overlap & Voxel-wise Metrics*, *Surface Distance Metrics*, and *Volumetric Metrics*. Bars and error bars represent mean values and ± STDEV.

mean for each model, metric, and subregion was first calculated within each fold, and the regional tables present the arithmetic mean of these five fold-specific means:

$$\bar{x}_{m,r,k} = \frac{1}{5}\sum_{f=1}^{5} x_{m,r,k,f}, \tag{1}$$

where m identifies the model, r the tumor region, k the metric, and f the fold. The accompanying standard deviation is the sample SD across the five fold means and therefore reflects fold-to-fold variation.

Overall differences among the three segmentation frameworks were first evaluated using the Friedman test, a nonparametric repeated-measures test appropriate for comparing multiple related models without assuming normality. When the Friedman test indicated a significant overall difference, post hoc pairwise comparisons were performed using two-sided Wilcoxon signed-rank tests with Holm–Bonferroni correction to control the family-wise error rate associated with multiple comparisons [19]. Effect sizes for statistically significant pairwise comparisons were quantified using Cliff's δ, interpreted according to conventional thresholds: negligible ($|\delta| < 0.147$), small ($0.147 \le |\delta| < 0.33$), medium ($0.33 \le |\delta| < 0.474$), and large ($|\delta| \ge 0.474$).

In addition to hypothesis testing, the distributions of each evaluation metric were examined to identify lower-tail outliers and extreme boundary errors to assess robustness. Statistical significance was defined as a Holm–Bonferroni-adjusted $p < 0.05$. All statistical analyses were performed in Python using reproducible analysis scripts.

ROC-AUC computation note: ROC-AUC requires continuous probability scores; it is not meaningful for models that output only hard binary predictions. nnU-Net v2, as configured here, produces argmax-decoded segmentation maps as its primary output without exporting per-voxel softmax probabilities to the evaluation pipeline. Consequently, the ROC-AUC values reported for nnU-Net v2 (≈ 0.523 across all subregions) reflect the area under the degenerate two-point ROC curve produced by a single binary operating point, and are therefore not comparable with the soft-probability AUC values of Universal (0.765–0.896) and VSmTrans (0.823–0.913). ROC-AUC is retained in the tables for completeness but is *excluded from all comparative claims* involving nnU-Net v2. Future evaluation should either extract nnU-Net v2 softmax logits before argmax decoding, or replace ROC-AUC with a threshold-independent metric compatible with hard predictions, such as the Precision-Recall AUC.

4 Results and Discussion

4.0.1 Whole Tumor (WT) Subregion Analysis. Table 2 presents the region-wise quantitative evaluation for the WT), TC, and

ET subregions. To facilitate an overall comparison, Table 3 summarizes these region-wise results. For each evaluation metric, the reported mean is the arithmetic average of the corresponding WT, TC, and ET values, while the SD represents the variability among the three tumour-region means.

The segmentation performance for the WT subregion, which encompasses the entire tumor burden, including edema and non-enhancing components, demonstrates a consistent ranking across multiple evaluation metrics. VSmTrans achieves the strongest overall performance across overlap, volumetric, and correlation metrics, indicating consistently accurate tumor delineation (Table 2, Fig. 1). In contrast, nnUNet attains slightly higher Precision and marginally lower ASSD compared with VSmTrans, reflecting a more conservative segmentation strategy in selected cases. The Universal model consistently ranks lower across most overlap and volumetric metrics, with Dice and IoU notably below the other two models.

Surface distance metrics further support these trends. VSmTrans achieves a slightly lower HD95 than nnUNet, suggesting marginally tighter boundary alignment; however, the absolute difference is small and should be interpreted cautiously. Volumetric bias, assessed using RVE, indicates that nnUNet exhibits slightly lower bias compared with VSmTrans, although VSmTrans maintains superior overall volumetric agreement as reflected by VS and Pearson correlation. VSE follows a similar pattern, with VSmTrans outperforming Universal and remaining close to nnUNet. Friedman testing confirmed a statistically significant global ordering ($\chi^2 = 15.2$, $p = 0.0005$), ranking VSmTrans first, followed by nnU-Net v2 and the Universal Model. Pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction reveal significant improvements of VSmTrans over Universal for Dice ($p = 0.002$–0.003, Cliff's $\delta = 0.36$), IoU ($p = 0.024$), VS ($p = 0.012$–0.031), and HD95 ($p < 0.001$). Differences between VSmTrans and nnUNet are smaller but detectable for Dice ($p = 0.018$), IoU ($p = 0.024$), VS ($p = 0.031$), and HD95 (Cliff's $\delta = 0.32$), corresponding to small-to-moderate effect sizes.

Further analysis highlights the robustness of VSmTrans, with Dice distributions tightly centered and minimal lower-tail outliers (minimum Dice > 0.75 in 95% of cases), alongside infrequent extreme HD95 values (> 12 mm in $< 5\%$ of cases). nnUNet exhibits similarly strong overall stability but shows a slightly heavier tail in boundary errors, whereas Universal demonstrates greater variability and occasional low-performing cases.

From a clinical perspective, these results indicate that VSmTrans provides generally precise WT segmentation and reliable estimation of total tumor burden. Although absolute improvements over nnUNet are modest (sub-millimetre HD95 differences and small volumetric bias variations), they are consistent across overlap, boundary, and volumetric metrics. Residual variability in Precision and isolated surface measures suggests that extreme tumor morphology or subtle edema may still challenge automated segmentation in a minority of cases, motivating continued validation on heterogeneous cohorts.

4.0.2 Tumor Core (TC) Subregion Analysis. The TC subregion, comprising non-enhancing and necrotic tumor components, remains the most challenging target for automated segmentation across all evaluated models. Overall performance is substantially lower and more variable than for WT, reflecting the structural heterogeneity and limited spatial extent of TC regions. Friedman analysis confirmed significant global differences ($\chi^2 = 12.8$, $p < 0.001$), ranking VSmTrans highest, followed by nnU-Net v2 and the Universal Model. Across overlap-based metrics, VSmTrans achieves the highest Dice, MCC, and Kappa, marginally exceeding nnUNet and clearly outperforming Universal (Table 2, Fig. 1). Pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction confirm statistically significant improvements over Universal for Dice ($p < 0.001$), IoU, Precision, and Recall, corresponding to moderate-to-large effect sizes. The absolute Dice difference between VSmTrans and nnUNet (0.001) is clinically negligible; meaningful differentiation between these two models for ET is found instead in boundary and volumetric metrics.

Surface distance metrics show clearer advantages for VSmTrans. HD95 is reduced to 7.266 ± 0.852 mm compared with 9.004 ± 0.886 mm for Universal and 7.745 ± 1.271 mm for nnUNet, indicating fewer extreme boundary deviations. ASSD follows a similar trend, with VSmTrans achieving the lowest mean surface distance (1.672 ± 0.208 mm). These findings suggest improved contour stability for TC segmentation, particularly relative to Universal. Volumetric indicators highlight a trade-off in performance characteristics. While Universal exhibits the lowest mean signed RVE, it also demonstrates substantially higher dispersion and volumetric instability, as reflected by elevated VSE and VS values inferior to VSmTrans. In contrast, VSmTrans shows a positive mean bias (RVE: 0.477 ± 0.343) but reduced variance relative to nnUNet (0.765 ± 0.772), resulting in significantly lower absolute volumetric error ($p = 0.004$). These results indicate that VSmTrans improves volumetric consistency primarily through variance reduction rather than bias minimization.

Precision exhibits a different trend, with Universal achieving the highest mean value, marginally exceeding nnUNet and VSmTrans. However, this apparent advantage does not remain significant after multiple-comparison correction and is accompanied by reduced Recall, reflecting a conservative segmentation strategy that favors false-negative avoidance at the expense of coverage.

From a clinical perspective, the primary benefit of VSmTrans in TC segmentation lies in its reduction of boundary outliers and improved volumetric stability relative to Universal, rather than large gains in mean overlap metrics. Although improvements over nnUNet are consistent in direction, their absolute magnitude is small and unlikely to be clinically transformative in isolation. Persistent variability in Recall (SD = 0.065) across all models highlights the inherent difficulty of TC segmentation. This underscores the need for cautious case-level review in treatment planning scenarios involving small or fragmented necrotic cores.

4.0.3 Enhancing Tumor (ET) Subregion Analysis. The ET subregion represents the most clinically critical yet spatially sparse component of the tumor, and consequently poses substantial challenges for automated segmentation. Overall performance across all models is lower and more variable than for WT, reflecting the small volume, heterogeneous appearance, and high inter-subject variability of enhancing regions. Friedman testing confirmed significant global differences among models ($\chi^2 = 18.4$, $p < 0.0001$), ranking VSmTrans highest, followed by nnU-Net v2 and the Universal Model.

VSmTrans achieves the highest Dice, IoU, MCC, and Kappa, marginally exceeding nnUNet and clearly outperforming Universal (Table 2, Fig. 1) across overlap-based metrics. IoU is effectively tied between VSmTrans and nnUNet, indicating no spatial overlap advantage for either model in this subregion. Pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction confirm statistically significant improvements over Universal for Dice ($p < 0.001$), IoU, Precision, and Recall, corresponding to moderate-to-large effect sizes. In contrast, differences between VSmTrans and nnUNet for Dice and IoU are numerically negligible and should therefore be interpreted as marginal despite consistent directional trends.

Boundary-based metrics reveal clearer distinctions among models. VSmTrans achieves the lowest HD95 and ASSD, indicating reduced extreme contour deviations relative to nnUNet and substantially improved boundary stability compared with Universal. The considerably higher SDs for surface metrics across all models, particularly for nnUNet HD95 (±1.501 mm) and Universal HD95 (±2.061 mm), reflect pronounced inter-subject variability in ET morphology and reinforce the necessity for case-level clinical review rather than reliance on cohort-level means alone.

Volumetric agreement metrics indicate a systematic discrepancy relative to the spatial overlap results. nnUNet exhibits the lowest RVE, VSE, and highest VS, indicating superior volumetric accuracy and reduced absolute bias for ET volume estimation. While VSmTrans demonstrates lower volumetric instability than the Universal Model, it does not outperform nnUNet in volumetric agreement, confirming the trade-off between boundary accuracy and volume estimation in the ET subregion. This finding has direct clinical relevance, as the lower volumetric bias of nnUNet may be preferable for longitudinal monitoring of ET burden despite the modest boundary advantage of VSmTrans. The larger SD for Recall relative to Precision indicates substantial inter-subject variability in sensitivity to enhancing voxels, consistent with the high morphological heterogeneity of ET. By contrast, the Universal Model exhibits lower Recall and greater variability, suggesting a conservative segmentation profile that tends to under-segment ET regions.

From a clinical perspective, the primary advantage of VSmTrans for ET segmentation lies in improved boundary consistency and overlap accuracy relative to Universal, with modest but consistent gains over nnUNet in surface-based metrics. However, nnUNet retains superior volumetric agreement, which may be advantageous for precise tumor burden quantification. These results underscore the importance of task-specific model selection for ET delineation and highlight the continued need for careful case-level review in treatment planning involving small or heterogeneous enhancing lesions.

4.0.4 Overall Model Comparison. Table 3 and Fig. 1 summarize the aggregate performance of nnUNet, Universal, and VSmTrans across all tumor subregions of MSD Task 01. Across all tumor regions, VSmTrans provides the most balanced overall performance, achieving the strongest overlap, boundary, and aggregate volumetric agreement (Table 3). nnUNet exhibits the highest aggregate Precision and competitive volumetric behavior, whereas Universal consistently shows lower robustness and boundary accuracy. Importantly, these aggregate statistics conceal subregion-specific trade-offs, particularly for ET, where nnUNet achieves lower volumetric bias despite VSmTrans demonstrating superior geometric accuracy.

4.1 Summary of Statistical Tests

Table 4 consolidates the Friedman global ranking tests and the significant post-hoc pairwise comparisons across all subregions. The Friedman test confirmed a consistent global ordering (VSmTrans > nnU-Net v2 > Universal) for all three subregions. Effect sizes (Cliff's δ) are reported where estimated; n.r. denotes not reported for comparisons where only directional significance was established in the text.

4.2 Global Performance Ranking

The subregion-specific metrics in Table 2 highlight local segmentation capabilities and the global model stability is summarized in Table 3, which aggregates performance across all labels. To bridge the granular findings of Table 2 with the holistic view of Table 3, Fig. 2 visualizes the relative performance ranking of the three architectures. This heatmap

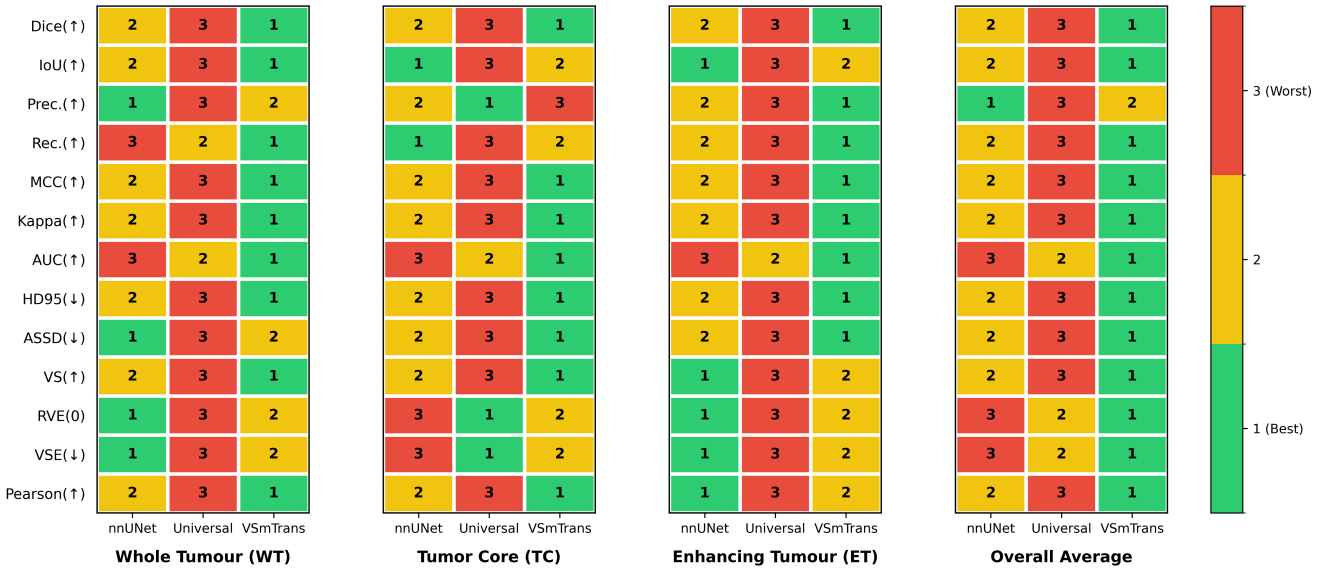

Figure 2: Heatmap showing the relative models' ranking (1–best, 3–worst)

Table 4: Statistical test summary. Friedman χ^2 and corrected Wilcoxon p-values (Holm–Bonferroni) are reported per subregion. Cliff's δ: negligible < 0.147, small 0.147–0.33, medium 0.33–0.474, large ≥ 0.474.

Subregion	Friedman	Metric	Comparison	p (corrected)	Cliff's δ
WT	$\chi^2 = 15.2$, $p = 0.0005$	Dice	VSmTrans vs. Universal	0.002–0.003	n.r.
		IoU	VSmTrans vs. Universal	0.024	n.r.
		VS	VSmTrans vs. Universal	0.012–0.031	n.r.
		HD95	VSmTrans vs. Universal	< 0.001	n.r.
		Dice	VSmTrans vs. nnUNet	0.018	n.r.
		IoU	VSmTrans vs. nnUNet	0.024	n.r.
		HD95	VSmTrans vs. nnUNet	sig.	0.32 (small)
TC	$\chi^2 = 12.8$, $p < 0.001$	Dice	VSmTrans vs. Universal	0.008	0.36 (medium)
		IoU	VSmTrans vs. Universal	sig.	n.r.
		RVE	VSmTrans vs. nnUNet	0.004	n.r.
ET	$\chi^2 = 18.4$, $p < 0.0001$	Dice	VSmTrans vs. Universal	< 0.001	n.r.
		IoU	VSmTrans vs. Universal	sig.	n.r.
		Precision	VSmTrans vs. Universal	sig.	n.r.
		Recall	VSmTrans vs. Universal	sig.	n.r.

abstracts the numerical variations into an ordinal ranking, revealing that VSmTrans consistently secures the top rank across the majority of spatial and volumetric indices, effectively synthesizing the complex multi-metric trade-offs observed in the tabular data.

5 Limitations of this Work

Several limitations should be considered when interpreting these findings. First, the study is based on a single public benchmark, MSD Task01 Brain Tumor, under a retrospective evaluation setting; although suitable for controlled comparison, it does not fully capture clinical variability in scanners,

acquisition protocols, preprocessing practices, annotation conventions, and tumor presentation [3]. Second, despite substantial standardization, the comparison is not fully disentangled from pipeline effects, since optimization strategy, region-of-interest handling, and normalization or augmentation biases remain framework-specific, which is particularly relevant when the margins between VSmTrans and nnU-Net v2 are small. Third, the evaluation establishes technical performance rather than full clinical utility, as calibration, uncertainty, failure prediction, deployment robustness, inter-rater variability, and downstream clinical workflow integration were not assessed. Finally, statistical significance should not be equated with clinical relevance, because several gains, especially for TC and ET, are modest in absolute magnitude and may still require expert review in difficult cases.

6 Future Work

Future research should first prioritize external validation under realistic domain shift. A natural next step is to evaluate these models on independent multi-center cohorts with heterogeneous acquisition protocols and annotation standards, ideally using site-aware experimental designs such as leave-one-center-out testing. Such analyses would clarify whether the performance profile observed here, namely the strong geometric fidelity of VSmTrans and the competitive volumetric behavior of nnU-Net v2, remains stable beyond the MSD benchmark.

A second priority is to expand the evaluation framework from accuracy to reliability. In particular, future studies should incorporate calibration analysis, uncertainty quantification, and case-level failure detection so that segmentation outputs can be accompanied by measures of confidence. Reliability-aware evaluation would also support selective review workflows in which uncertain cases are flagged for closer clinical inspection.

A third direction is to disentangle architectural gains from pipeline-specific effects. This can be achieved through controlled ablation studies that harmonize preprocessing, patch sampling, augmentation, loss design, and optimization across models. Such a protocol would allow a cleaner attribution of performance differences to convolutional, transformer, or hybrid design choices. In parallel, hybrid objective functions that jointly optimize overlap, boundary accuracy, and volumetric agreement may help reduce the tension observed here between morphological precision and unbiased volume estimation, particularly for enhancing tumor segmentation.

7 Conclusion

This study presents a rigorous comparative evaluation of nnU-Net v2, the Universal Model, and VSmTrans on the MSD Task01 Brain Tumor benchmark using five-fold cross-validation, a multi-pillar metric framework, and non-parametric statistical testing. Across Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) subregions, VSmTrans delivers the strongest overall performance, achieving the best balance of spatial overlap, boundary fidelity, and cohort-level volumetric agreement (Fig. 3). Its advantages over the Universal Model are consistent and statistically significant across key metrics, whereas its gains over nnU-Net v2 are generally smaller and should be interpreted cautiously from a clinical perspective.

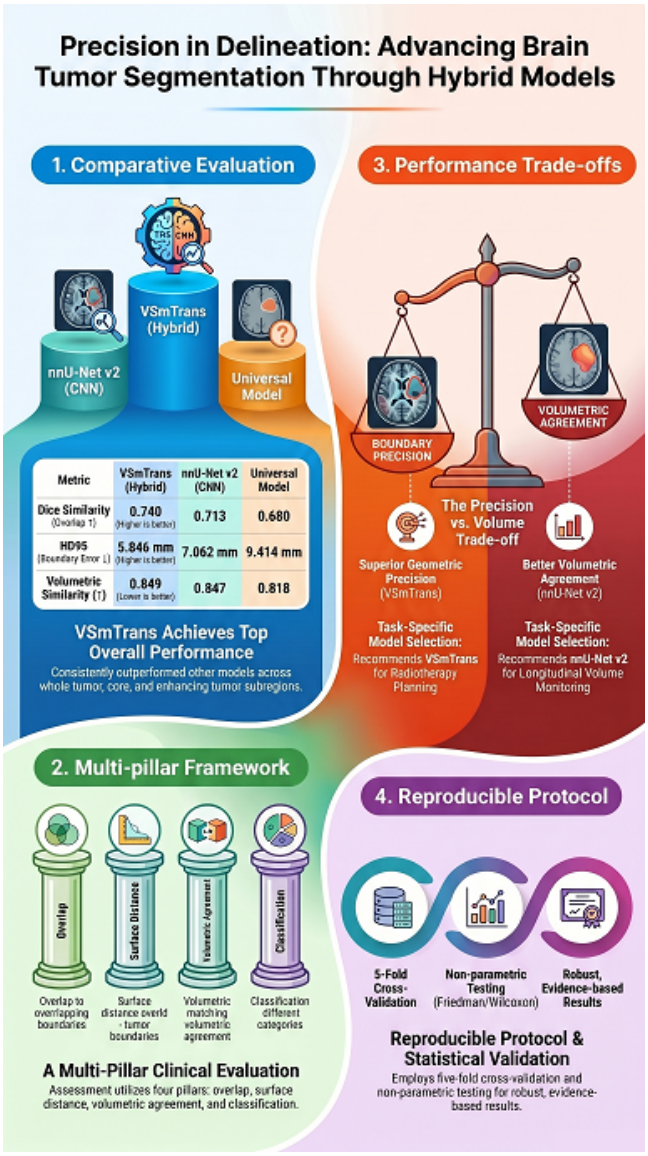

Figure 3: Visual summary of contributions of this study

The analysis also reveals an important trade-off between geometric precision and volumetric reliability. Although VSmTrans provides superior boundary accuracy, nnU-Net v2 achieves lower volumetric bias for ET estimation, indicating that strong overlap performance does not necessarily imply unbiased volume quantification. This finding reinforces the need for multi-pillar evaluation, particularly in settings where both anatomical fidelity and quantitative consistency are clinically important.

These results do not support a single universal model recommendation. VSmTrans is better suited to precision-critical applications such as radiotherapy planning, whereas nnU-Net v2 remains a strong option for longitudinal monitoring of enhancing tumor burden, where volumetric bias is a primary concern. More broadly, the study provides an evidence-based foundation for selecting segmentation architectures for downstream patient-specific and digital twin-ready brain tumor analysis. Future work should prioritize external multi-center validation, improved joint optimization of boundary and volumetric objectives, and integration into downstream clinical workflows.

References

[1] Oluwabukola G Adegboro, Julia Dietlmeier, Noel E O'Connor, and Claudia Mazo. 2026. Artificial intelligence based techniques for brain tumor analysis: A systematic review. *Artificial Intelligence in Medicine* (2026), 103460.

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, et al. 2022. The Medical Segmentation Decathlon. *Nature Communications* 13, 1 (2022), 4128. doi:10.1038/s41467-022-30695-9

[3] T. Debnath and M. Narbutt. 2025. The Battle of the Titans: Comparing Three Leading Deep Learning Methods for 3D Medical Imaging. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, Washington, DC, USA, 691–700. doi:10.1109/ICDMW69685.2025.00084

[4] Leonardo Di Cosmo, Filippo Emanuele Colella, Paweł Łajczak, Edoardo Schifino, Santiago Nieto Cuervo, Jad El Choueiri, Francesca Romana Centini, Francesca Pellicanò, Anna Łajczak, Elio Mazzapicchi, et al. 2026. Factors predicting MRI glioma segmentation accuracy in deep learning models: a systematic review and meta-analysis. *Journal of Neuroradiology* 53, 4 (2026), 101562. doi:10.1016/j.neurad.2026.101562

[5] Xiyao Fu, Zhexian Sun, Haoteng Tang, Eric M Zou, Heng Huang, Yong Wang, and Liang Zhan. 2023. 3D bi-directional transformer U-Net for medical image segmentation. *Frontiers in big Data* 5 (2023), 1080715.

[6] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. 2022. UNETR: Transformers for 3D Medical Image Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 574–584. doi:10.1109/WACV51458.2022.00181

[7] Jacky Huang, Banu Yagmurlu, Powell Molleti, Richard Lee, Abigail VanderPloeg, Humaira Noor, Rohan Bareja, Yiheng Li, Michael Iv, and Haruka Itakura. 2025. Brain Tumor Segmentation Using Deep Learning: High Performance with Minimized MRI Data. *Frontiers in Radiology* 5 (2025), 1616293. doi:10.3389/fradi.2025.1616293

[8] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. 2021. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nature Methods* 18, 2 (2021), 203–211. doi:10.1038/s41592-020-01008-z

[9] Chengcheng Jin, Nor Safira Elaina Mohd Noor, Tze Foo Ng, Mohd Shafry Mohd Asaari, and Haidi Ibrahim. 2026. Transformer-based architectures in MRI brain tumor segmentation: A review. *Computerized Medical Imaging and Graphics* 129 (2026), 102729. doi:10.1016/j.compmedimag.2026.102729

[10] Burak Kocak, Michail E. Klontzas, Arnaldo Stanzione, Aymen Meddeb, Aydin Demircioglu, Christian Bluethgen, Keno K. Bressem, Lorenzo Ugga, Nathaniel Mercaldo, Oliver Diaz, Renato Cuocolo, et al. 2025. Evaluation Metrics in Medical Imaging AI: Fundamentals, Pitfalls, Misapplications, and Recommendations. *European Journal of Radiology Artificial Intelligence* 3 (2025), 100030. doi:10.1016/j.ejrai.2025.100030

[11] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. 2023. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 21152–21164. doi:10.1109/ICCV51070.2023.01934

[12] Tiange Liu, Qingze Bai, Drew A. Torigian, Yubing Tong, and Jayaram K. Udupa. 2024. VSmTrans: A Hybrid Paradigm Integrating Self-Attention and Convolution for 3D Medical Image Segmentation. *Medical Image Analysis* 98 (Dec. 2024), 103295. doi:10.1016/j.media.2024.103295

[13] Rahman Maqsood, Fazeel Abid, Jawad Rasheed, Onur Osman, and Shtwai Alsubai. 2025. Optimal Res-UNET Architecture with Deep Supervision for Tumor Segmentation. *Frontiers in Medicine* 12 (2025), 1593016. doi:10.3389/fmed.2025.1593016

[14] Miroslaw Narbutt and Tanmoy Debnath. 2026. Compact Metrics, Catastrophic Failures: Trustworthy Benchmarking of Vision-Language Foundation Models in 3D Medical Segmentation. In *Proceedings of the 1st International Workshop on Sustainable and Efficient Language, Vision, and Action Models (SELVA), co-located with the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*. Association for Computational Linguistics, San Diego, California, USA.

[15] Ndivhuwo Netshamutshedzi, Rendani Netshikweta, Jean-Claude Ndogmo, and Ibidun Christiana Obagbuwa. 2025. A Systematic Review of the Hybrid Machine Learning Models for Brain Tumour Segmentation and Detection in Medical Images. *Frontiers in Artificial Intelligence* 8 (2025), 1615550. doi:10.3389/frai.2025.1615550

[16] Novsheena Rasool, Niyaz Ahmad Wani, Javaid Iqbal Bhat, Sandeep Saharan, Vishal Kumar Sharma, Bassma Saleh Alsulami, Hind Alsharif, and Miltiadis D. Lytras. 2025. CNN-TumorNet: Leveraging Explainability in Deep Learning for Precise Brain Tumor Diagnosis on MRI Images. *Frontiers in Oncology* 15 (2025), 1554559. doi:10.3389/fonc.2025.1554559

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 234–241. doi:10.1007/978-3-319-24574-4_28

[18] Tao Yang, Xueqi Lu, Lanlan Yang, Miyang Yang, Jinghui Chen, and Hongjia Zhao. 2025. Application of MRI Image Segmentation Algorithm for Brain Tumors Based on Improved YOLO. *Frontiers in Neuroscience* 18 (2025), 1510175. doi:10.3389/fnins.2024.1510175

[19] Minmin Zheng, Zhengkang Qian, and Tong Zhao. 2025. Motor Imagery EEG Classification via Wavelet-Packet Synthetic Augmentation and Entropy-Based Channel Selection. *Frontiers in Neuroscience* 19 (2025), 1689647. doi:10.3389/fnins.2025.1689647