Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

Region-Aware Boundary-Weighted Dice Does Not Resolve the Core–Enhancing Trade-off in Brain Tumor Segmentation: A Controlled Negative Result on BraTS 2020

S. U. Ravi Kumar Chavali[1], P. V. Kumar[2]

[1]Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India
[2]School of Engineering, Anurag University, Hyderabad, Telangana, India

Abstract:
A companion matched-condition study found that a boundary-weighted Dice loss (BW-Dice) improves tumor-core (TC) segmentation but degrades enhancing-tumor (ET) segmentation — a trade-off with no net gain — and hypothesized a scale-mismatch mechanism in which the boundary-emphasis kernel is wider than the small ET region. The natural prediction of that hypothesis is tested: shielding ET from the boundary emphasis should preserve the TC gain while removing the ET loss. Under matched conditions on BraTS 2020, a region-aware BW-Dice variant is trained that applies boundary weighting to the necrotic-core and edema channels but trains ET with standard Dice, evaluated across three random seeds, a kernel-width sweep, and two additional decoders. Region-aware weighting does not improve on BW-Dice: in numerically stable settings it is equal to or worse than BW-Dice on both TC and ET, and a single-seed apparent gain proves non-reproducible. The core–enhancing trade-off is therefore not a fixable artifact of kernel scale. Seed- and architecture-dependent mixed-precision instability of weighted-Dice losses is also documented. Reporting this negative result narrows the space of promising loss-design interventions.

Keywords: Brain Tumor Segmentation, BraTS 2020, Boundary-Weighted Loss, Negative Result, Reproducibility

1. Introduction

1.1 Clinical Context and the Core-Enhancing Trade-off

Automated segmentation of gliomas from multi-modal magnetic resonance imaging supports diagnosis, surgical planning, and radiotherapy target definition. Following the Brain Tumor Segmentation (BraTS) convention, accuracy is reported on three hierarchical sub-regions: the whole tumor (WT), the tumor core (TC), and the enhancing tumor (ET). These sub-regions differ markedly in size and difficulty: WT is large and comparatively easy, whereas ET is small, thin, and the most sensitive to segmentation error. A

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

recurring difficulty in loss-function design for this task is that interventions which help one sub-region often harm another.

A companion study [1] examined exactly such an intervention. It introduced a boundary-weighted Dice loss (BW-Dice) that up-weights voxels near the necrotic-core-edema interface, a region where the standard Dice loss tends to under-segment. BW-Dice improved TC Dice by a small but consistent margin, but it degraded ET Dice by a comparable amount, so that the aggregate mean Dice was unchanged. That study reported the effect honestly as a trade-off rather than an improvement, and proposed a mechanistic explanation: because the Gaussian boundary-emphasis kernel has a spatial width comparable to or larger than the ET region itself, the weighting cannot create a meaningful interior-versus-boundary distinction for ET and instead diverts optimization pressure toward the larger necrotic-core-edema boundary, at the expense of ET.

1.2 Hypothesis: Is the Trade-off a Fixable Artifact

The scale-mismatch explanation of [1] makes a testable prediction. If ET degrades specifically because the emphasis kernel is too wide relative to ET, then shielding ET from the boundary emphasis — while retaining that emphasis for the necrotic-core and edema channels that drive the TC gain — should remove the ET loss while preserving the TC benefit, yielding a net improvement. This intervention is called region-aware BW-Dice. This paper tests that prediction under matched conditions, with multi-seed replication, a kernel-width sweep, and a cross-architecture check.

The result is negative. Region-aware BW-Dice does not improve on BW-Dice. In numerically stable configurations it is equal to or worse than BW-Dice on both TC and ET, and a promising single-seed pilot result proves not to reproduce across random seeds. This outcome is reported because a well-designed negative result is informative: it rules out a natural and otherwise plausible hypothesis, and it re-confirms, on a second independent example, the single-seed evaluation hazard that motivated multi-seed testing in [1].

1.3 Contributions

C1 — A controlled test of a mechanism-driven hypothesis. The prediction that protecting ET from boundary emphasis should convert the BW-Dice trade-off into a net gain is derived from the scale-mismatch explanation of [1], and is tested directly with a matched-condition experiment.

C2 — A clean negative result. Across three seeds and a kernel-width sweep, region-aware BW-Dice does not beat BW-Dice; the hypothesis is not supported, indicating that the core-enhancing trade-off is not merely an artifact of boundary-kernel scale.

C3 — A reproducibility cautionary tale. A single-seed pilot showed an apparent improvement that vanished under replication, with one of three seeds collapsing entirely; this is quantified to illustrate the danger of drawing conclusions from single training runs.

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

C4 — A documented instability of weighted-Dice losses. Seed- and architecture-dependent numerical instability of custom weighted-Dice losses under mixed-precision training is observed, a practical caution for practitioners implementing such losses.

2. Background and Related Work

2.1 Boundary- and Distance-Weighted Losses

Spatially weighted losses have a long history in segmentation. The original U-Net [5] introduced a pre-computed pixel weight map that emphasizes thin separations between touching objects. Distance-transform and boundary losses [8] similarly bias optimization toward accurate boundary placement, and have been reported to help on structures with elongated or intricate borders. The BW-Dice loss of [1] belongs to this family: it multiplies each contribution to the Dice overlap by a weight that peaks at a chosen class boundary and decays with a Gaussian profile of width controlled by a parameter. Such methods implicitly assume that the emphasized structures are large relative to the emphasis kernel; the present work probes what happens when that assumption fails for a small sub-region.

2.2 The Companion BW-Dice Study

The immediate predecessor to this work [1] is a matched-condition evaluation of BW-Dice against standard Dice on BraTS 2020. Its central finding was a regional trade-off: TC improved, ET degraded, and mean Dice was statistically unchanged across seeds. That study attributed the ET degradation to a scale mismatch between the emphasis kernel and the ET region, and explicitly flagged a region-aware remedy as future work. The present paper implements and tests that remedy. It also inherits the methodological lesson of the earlier study, learned when a single-seed configuration appeared to win but did not survive multi-seed replication.

2.3 Negative Results and Reproducible Evaluation

The reliability of small accuracy gains in segmentation has come under increasing scrutiny. Rigorous re-evaluations [12] have shown that many reported architectural improvements fail to survive matched, well-controlled comparison, and analyses of benchmark variance [15] show that gains within the range of seed-to-seed fluctuation are not trustworthy without replication. Reporting negative results in this setting is valuable precisely because it prevents the literature from accumulating irreproducible micro-improvements. The contribution here is in that spirit: a carefully controlled test that returns a clear negative, together with the evidence needed to trust it.

3. Method

3.1 Data and Matched Protocol

All experiments use the BraTS 2020 training set (369 annotated cases) under the identical preprocessing and split used in [1]: a fixed 80/10/10 case-level partition, 2D axial slices containing at least a minimal tumor area, per-modality z-score normalization, a centre crop resized to 128 × 128, and four input channels (T1, T1ce, T2, FLAIR). Every model is a decoder from the segmentation-models-pytorch library with a

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

shared ResNet-34 encoder pre-trained on ImageNet. Training uses the Adam optimizer at a learning rate of 1×10^{-4} with cosine annealing, a maximum of 25 epochs with early stopping (patience 5) on validation mean Dice, mixed-precision (FP16) computation, and horizontal and vertical flip augmentation. Holding all of this fixed isolates the effect of the loss function.

3.2 Boundary-Weighted Dice

Let the ground-truth label map define, for each pixel p, its Euclidean distance d(p) to the nearest class-to-class boundary. The BW-Dice weight map of [1] is given in Equation (1),

$$w(p) = 1 + \lambda \cdot m_k(p) \cdot \exp\left(-\frac{d(p)^2}{2\sigma^2}\right) \quad (1)$$

where λ controls emphasis strength, σ controls the Gaussian width, and $m_k(p)$ is a class-specific multiplier equal to a value m_{NE} when the nearest boundary is of the necrotic-core-edema type and 1 otherwise. Throughout, the setting reported best in [1] is used ($\lambda = 2$, $m_{NE} = 2$), and σ is varied. The weight map enters a soft multiclass Dice loss over the foreground classes C = {necrotic core, edema, ET}, given in Equation (2),

$$L_{BW} = 1 - \frac{1}{|C|} \sum_{c \in C} \frac{2 \Sigma w(p) P_c(p) G_c(p)}{\Sigma w(p) P_c(p) + \Sigma w(p) G_c(p)} \quad (2)$$

where $P_c(p)$ is the predicted probability of class c at pixel p, $G_c(p)$ the corresponding ground-truth indicator, and the inner sums run over pixels. Setting $\lambda = 0$ (equivalently $w \equiv 1$) recovers standard Dice; this same loss with weighting disabled is used as the Dice baseline, so that the baseline and the weighted variants differ only in the weighting and are otherwise byte-for-byte identical in reduction, smoothing, and class handling.

3.3 The Region-Aware Intervention

The region-aware variant implements the remedy predicted by the scale-mismatch hypothesis. It replaces the shared weight map with a per-class weight $w_c(p)$: the necrotic-core and edema channels receive the full boundary weighting w(p), preserving the emphasis that produced the TC gain in [1], while the ET channel is fixed to unit weight and therefore trained with standard, unweighted Dice. If ET degrades because the emphasis kernel is too wide relative to ET, protecting the ET channel in this way should eliminate the degradation while leaving the TC benefit intact.

3.4 Experimental Design

Three complementary experiments test the hypothesis. First, a core multi-seed comparison trains UNet++ with each of the three losses — Dice, BW-Dice, and region-aware BW-Dice — at $\sigma = 5$ across three random seeds (42, 123, 2024), reporting mean and standard deviation and per-seed sign consistency.

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

Second, a kernel-width sweep repeats the BW-Dice and region-aware conditions at $\sigma \in \{3, 5, 7\}$, three seeds each, to test the mechanistic prediction that protecting ET should matter more as the emphasis band widens. Third, a cross-architecture check repeats the three losses on U-Net and MAnet at $\sigma = 5$. In total, 27 models are trained under identical conditions.

3.5 Evaluation

Models are evaluated by per-slice Dice on the held-out test slices for each of WT, TC, and ET, with a slice in which a region is absent from both prediction and ground truth scored as 1. This per-slice protocol matches [1] exactly, keeping the two studies directly comparable. Where a significance statement is made, it uses a paired bootstrap over slices with 10,000 resamples. It is noted, as in [1], that slice-level resampling is mildly anti-conservative because slices from one patient are correlated; this does not affect the negative conclusion, which turns on effects being absent or wrong-signed rather than on marginal significance.

4. Results

4.1 Region-Aware Weighting Does Not Beat BW-Dice

Table 1 reports the core comparison on UNet++ at $\sigma = 5$, as per-slice Dice averaged over seeds 42, 123, and 2024. Averaged over three seeds, region-aware BW-Dice attains the lowest mean Dice of the three losses (0.789), below both BW-Dice (0.824) and the standard Dice baseline (0.817). Far from protecting ET, the region-aware variant records the lowest ET of the three (0.762) and the lowest TC (0.756). Its large standard deviations on TC and ET (±0.076 and ±0.072) are the visible signature of a single unstable seed, examined next. The plain BW-Dice column reproduces the finding of [1]: relative to Dice it raises TC (0.806 versus 0.801) and, in this run, holds ET essentially level; it is BW-Dice, not the region-aware variant, that is the strongest loss here.

Table 1: Core Comparison on UNet++ at $\sigma = 5$ (Per-Slice Dice, Mean ± SD Over Seeds 42, 123, 2024)

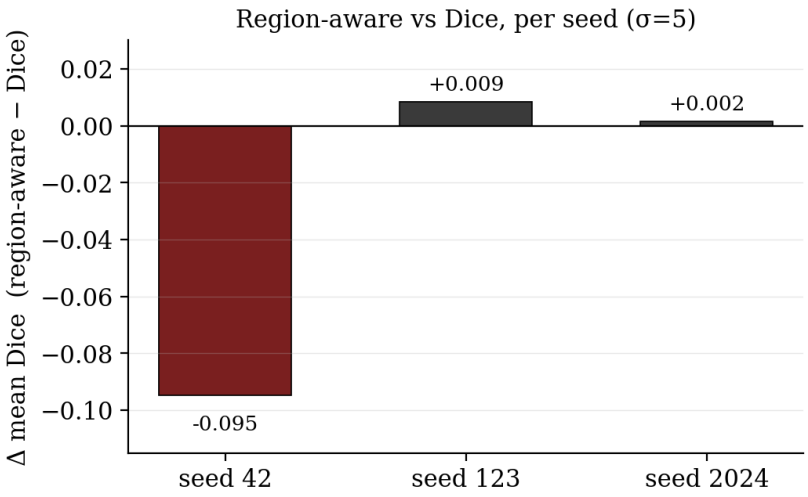
Loss variant	WT	TC	ET	Mean
Standard Dice	0.841 ± 0.010	0.801 ± 0.012	0.809 ± 0.005	0.817 ± 0.008
BW-Dice [1]	0.851 ± 0.003	0.806 ± 0.009	0.816 ± 0.001	0.824 ± 0.003
Region-aware BW-Dice	0.849 ± 0.003	0.756 ± 0.076	0.762 ± 0.072	0.789 ± 0.050

4.2 The Single-Seed Pilot Result Does Not Reproduce

A single-seed pilot of the region-aware variant at $\sigma = 5$ had appeared promising, reaching a mean Dice comparable to BW-Dice. Replication dissolves that impression. The per-seed change in mean Dice relative to the matched Dice baseline is +0.008 (seed 123) and +0.002 (seed 2024) — both within seed-to-seed noise — but −0.095 (seed 42), a collapse in which the model failed to train stably under the region-aware loss. Figure 1 shows the three seeds together.

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

Figure 1: Per-Seed Change in Mean Dice for Region-Aware BW-Dice Relative to the Standard-Dice Baseline (σ = 5). One Seed Collapses; the Single-Seed Pilot Corresponded to a Favorable Draw, Not a Reproducible Effect.



This is the same failure mode documented in [1], where a single-seed configuration appeared to win and then evaporated under replication. It is the reason the present study was designed with three seeds from the outset, and it is the strongest practical lesson of the paper: a single training run is not evidence of an effect in this regime.

4.3 The Kernel-Width Sweep Does Not Support the Mechanism

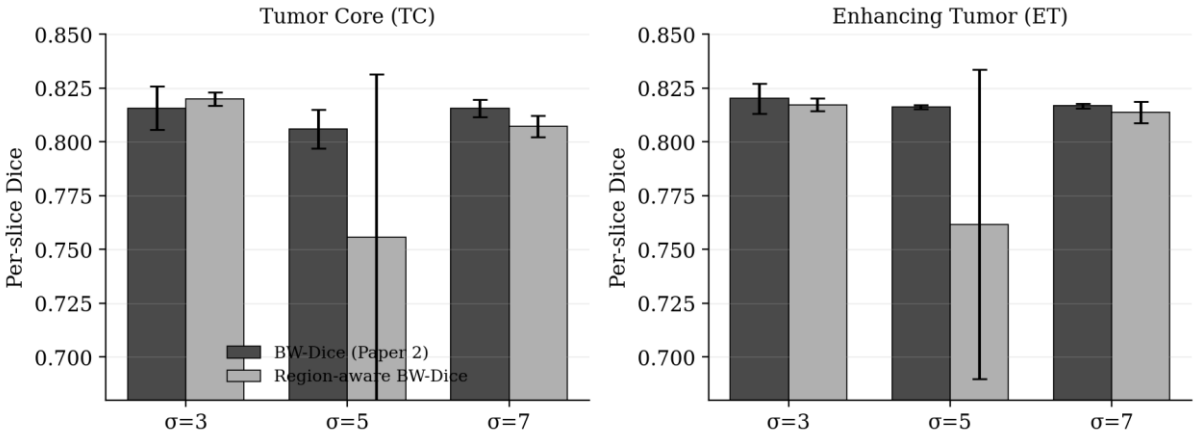
Table 2 reports the kernel-width sweep on UNet++, giving TC and ET per-slice Dice for plain BW-Dice and region-aware BW-Dice at $\sigma \in \{3, 5, 7\}$. In the numerically stable rows ($\sigma = 3$ and $\sigma = 7$), region-aware is equal to or worse than BW-Dice on both regions; the $\sigma = 5$ region-aware row is contaminated by the unstable seed of Section 4.2. The same picture is shown in Figure 2.

Table 2: Kernel-Width Sweep on UNet++ (Per-Slice TC and ET Dice, Mean ± SD Over Three Seeds)

σ	Variant	TC	ET
3	BW-Dice	0.816 ± 0.010	0.820 ± 0.007
3	Region-aware	0.820 ± 0.003	0.817 ± 0.003
5	BW-Dice	0.806 ± 0.009	0.816 ± 0.001
5	Region-aware	0.756 ± 0.076	0.762 ± 0.072
7	BW-Dice	0.816 ± 0.004	0.817 ± 0.001
7	Region-aware	0.807 ± 0.005	0.814 ± 0.005

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

Figure 2: Kernel-Width Sweep — Tumor-Core (Left) and Enhancing-Tumor (Right) Per-Slice Dice for BW-Dice Versus Region-Aware BW-Dice at σ ∈ {3, 5, 7} (Mean ± SD Over Three Seeds). Region-Aware Never Exceeds BW-Dice on ET; the Tall Error Bar at σ = 5 Is the Unstable Seed.



The mechanistic prediction was specific: if the ET degradation is caused by an over-wide emphasis kernel, then protecting ET should help increasingly as σ grows, and the ET of plain BW-Dice should fall as σ grows. Neither pattern appears. The ET of plain BW-Dice is essentially flat across σ (0.820, 0.816, 0.817), so there is no widening-driven degradation for the region-aware variant to rescue. And in the two numerically stable kernel widths, σ = 3 and σ = 7, region-aware weighting is equal to or slightly worse than plain BW-Dice on both TC and ET. Protecting the ET channel simply does not recover ET, and it does not improve on BW-Dice. The scale-mismatch hypothesis, as an explanation that a region-aware remedy could exploit, is not supported.

4.4 Cross-Architecture and Loss-Stability Observations

Table 3 reports the cross-architecture check at σ = 5, seed 42. The check is inconclusive by accident rather than by design, and is reported plainly. Two baseline configurations collapsed — U-Net with standard Dice to a mean of 0.198, and MAnet with BW-Dice to 0.674 — which are failures of those baselines, not of the region-aware variant, and preclude a fair cross-architecture comparison from this single-seed run. The collapses are consistent with the mixed-precision fragility of custom Dice-family losses noted in [1]. On the configurations that did train, region-aware weighting again fails to exceed plain BW-Dice (U-Net: 0.809 versus 0.828). No generalization claim is drawn from Table 3 beyond noting the instability itself, which is a practical caution for anyone implementing weighted-Dice losses under mixed precision.

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

Table 3: Cross-Architecture Check at σ = 5, Seed 42 (Per-Slice Dice; Two Baseline Rows Collapsed During Training)

Architecture	Variant	WT	TC	ET	Mean
U-Net	Dice	0.079	0.027	0.486	0.198
U-Net	BW-Dice	0.859	0.810	0.816	0.828
U-Net	Region-aware	0.832	0.788	0.806	0.809
MAnet	Dice	0.809	0.748	0.756	0.771
MAnet	BW-Dice	0.713	0.618	0.690	0.674
MAnet	Region-aware	0.832	0.781	0.790	0.801

5. Discussion

5.1 The Trade-off Is Not a Kernel-Scale Artifact

The clearest reading of these results is that the core-enhancing trade-off reported in [1] is not an artifact of the boundary-emphasis kernel being too wide for ET. Had it been, protecting the ET channel would have recovered ET; it did not. The stable kernel widths show region-aware weighting matching or trailing plain BW-Dice on ET rather than improving it, and plain BW-Dice shows no σ-driven ET degradation for a remedy to undo. The tension between tumor-core and enhancing-tumor accuracy therefore appears to be more fundamental — rooted in the shared representation and the competition among classes for capacity — than a spatial re-weighting of the loss can resolve. This narrows the search: future attempts to break the trade-off should look to architectural or multi-task mechanisms, for example region-decoupled decoders or auxiliary ET-specific objectives, rather than to further refinements of boundary weighting.

5.2 A Second Lesson in Single-Seed Evaluation

The most transferable finding is methodological. A single seed of the region-aware variant looked like a success and would, on its own, have supported a positive-result paper. Two additional seeds revealed that the apparent gain was a favorable draw and that the configuration is in fact unstable. This is the second time in this line of work that single-seed evaluation produced a mirage, and it is a concrete instance of the benchmark-variance concerns raised in the wider literature [15]. For accuracy differences on the order of a Dice point — the scale of most reported loss-design gains — replication across seeds is not optional.

5.3 Mixed-Precision Fragility of Weighted Losses

The collapses in Tables 1 and 3 point to a practical hazard. Custom weighted-Dice losses, especially when combined with certain decoders under FP16 mixed precision, can train unstably in a seed- and architecture-dependent way. Practitioners adopting such losses should monitor for this, validate across seeds, and

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

consider full-precision training or loss re-formulation where instability appears. The instability is reported rather than silently discarded, because concealing it would misrepresent the reliability of the method.

5.4 Limitations

This study shares the scope of its predecessor: 2D axial slices, a single dataset (BraTS 2020), and a per-slice evaluation whose slice-level resampling is mildly anti-conservative. These limits do not weaken a negative conclusion — an effect that is absent or wrong-signed under these conditions is not rescued by a more favorable evaluation — but they do bound the generality of the claim. One specific region-aware formulation is tested (unit weight on the ET channel); other formulations, such as a per-class kernel width, are conceivable, though the flat σ-response of plain BW-Dice gives little reason to expect them to succeed where this one failed.

6. Conclusion

A natural, mechanism-driven hypothesis was tested: that the tumor-core and enhancing-tumor trade-off induced by boundary-weighted Dice could be resolved by shielding the enhancing tumor from the boundary emphasis. Under matched conditions on BraTS 2020, across three seeds, a kernel-width sweep, and two additional decoders, the region-aware variant did not improve on boundary-weighted Dice; in stable settings it was equal to or worse on both regions, and an apparent single-seed gain did not reproduce. The trade-off is therefore not a fixable artifact of boundary-kernel scale, and boundary re-weighting is unlikely to be the route to breaking it. Alongside this negative result, it was documented, for a second time in this line of work, how a single-seed run can manufacture a positive impression that replication removes, and a mixed-precision instability of weighted-Dice losses was flagged. Reporting these findings keeps the surrounding literature honest and redirects effort toward architectural rather than loss-weighting solutions.

7. Acknowledgement

The authors thank the organizers of the BraTS challenge for the dataset. Computations used the Google Colaboratory GPU service. This work received no specific funding. All experiments reuse the matched preprocessing pipeline of the companion study, and code for the region-aware loss and the full experiment matrix is available from the authors on request.

REFERENCES:

1. S. U. R. K. Chavali and P. V. Kumar, "Boundary-Weighted Dice Loss for Brain Tumor Segmentation: A Controlled Study of a Region-Specific Accuracy Trade-off on BraTS 2020," manuscript, 2024 (companion paper).
2. S. U. R. K. Chavali and P. V. Kumar, "A Controlled Comparison of U-Net Decoder Architectures and Loss Functions for Brain Tumor Segmentation on BraTS 2020," manuscript, 2024.
3. B. H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.

Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487
editor@aijmr.com
Volume 4, Issue 4, July-August 2026
CrossRef DOI: 10.62127/aijmr.2026.v04i04.1463

4. S. Bakas et al., "Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features," Scientific Data, vol. 4, art. 170117, 2017.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, pp. 234–241, 2015.
6. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in Proc. 3D Vision (3DV), pp. 565–571, 2016.
7. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in DLMIA, LNCS 11045, pp. 3–11, 2018.
8. H. Kervadec et al., "Boundary Loss for Highly Unbalanced Segmentation," in Proc. MIDL, pp. 285–296, 2019.
9. S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," in MLMI, LNCS 10541, pp. 379–387, 2017.
10. N. Abraham and N. M. Khan, "A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation," in Proc. IEEE ISBI, pp. 683–687, 2019.
11. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.
12. F. Isensee et al., "nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation," in Proc. MICCAI, arXiv:2404.09556, 2024.
13. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles," in Proc. NeurIPS, pp. 6402–6413, 2017.
14. P. Micikevicius et al., "Mixed Precision Training," in Proc. ICLR, arXiv:1710.03740, 2018.
15. X. Bouthillier et al., "Accounting for Variance in Machine Learning Benchmarks," in Proc. MLSys, 2021.
16. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation," Neurocomputing, vol. 338, pp. 34–45, 2019.