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.v04i04.1471

Test-Time Augmentation Consistently Improves U-Net-Family Brain Tumor Segmentation: A Controlled Study 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:
Test-time augmentation (TTA) averages a model's predictions over label-preserving input transformations and is a common, training-free addition to segmentation pipelines. Its isolated effect, however, is rarely quantified under matched conditions across a family of models. In this work, TTA is evaluated as a standalone intervention on sixteen previously trained U-Net-family configurations for brain tumor segmentation on BraTS 2020 (four decoders × four losses, sharing a ResNet34 encoder). Each model is evaluated on 2,242 test slices with and without four-flip TTA, and the difference is assessed by paired bootstrap. TTA improves mean per-slice Dice for every one of the sixteen models, with all sixteen improvements statistically significant. Pooled across models, TTA raises mean Dice from 0.813 to 0.826 (Δ = +0.0129, 95% CI [+0.0119, +0.0140], $p < 0.001$). The gain is largest for the architectures that were weakest without augmentation, so that TTA also narrows the performance gap between decoder families. Because TTA requires no additional training and only a fixed multiple of inference cost, it is a reliable and low-cost improvement for U-Net-family brain tumor segmentation.

Keywords: Test-Time Augmentation, Brain Tumor Segmentation, BraTS 2020, U-Net, Dice Score, Medical Image Segmentation

1. Introduction

1.1 Background and Motivation

Automated segmentation of gliomas from multi-modal magnetic resonance imaging (MRI) 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), measured by the Dice similarity coefficient. Deep convolutional networks of the U-Net family dominate this task, and a large literature proposes architectural and loss-function refinements aimed at small accuracy gains.

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.v04i04.1471

Test-time augmentation (TTA) is an inference-time technique that averages a model's predictions over a set of label-preserving transformations of the input, such as flips. It requires no additional training and is widely reported to improve both accuracy and uncertainty estimation [8]. In practice TTA is often bundled with other components of a pipeline, so its isolated contribution across a controlled family of models is seldom measured. This work quantifies that contribution directly.

A prior matched-condition study [1] trained sixteen U-Net-family configurations under strictly identical conditions and found that no single architecture or loss dominates on aggregate accuracy. Those sixteen models form an ideal controlled testbed for isolating the effect of TTA: because every model was trained identically, any difference produced by adding TTA at inference is attributable to the augmentation alone, not to differences in preprocessing, augmentation during training, or optimization.

1.2 Contributions

C1 — TTA improves every model, significantly. Across all sixteen matched U-Net-family configurations, four-flip TTA improves mean per-slice Dice, and in every case the improvement is statistically significant by paired bootstrap.

C2 — A consistent pooled gain. Pooled across all sixteen models, TTA raises mean Dice from 0.813 to 0.826 ($\Delta = +0.0129$, $p < 0.001$), a larger and more consistent effect than model ensembling reported on the same models.

C3 — TTA narrows the architecture gap. The gain is largest for the decoder families that were weakest without augmentation, so TTA reduces the spread between architectures rather than uniformly shifting all of them.

2. Related Work

Test-time augmentation has been studied as a means of improving both accuracy and calibrated uncertainty in medical image segmentation. Wang et al. [8] formalized TTA for segmentation and showed that averaging predictions over input transformations reduces aleatoric error and yields useful uncertainty estimates. TTA is a routine component of high-performing BraTS pipelines, including the nnU-Net framework [6, 7], where flip-based augmentation at inference is applied as a standard step. Because TTA operates by averaging predictions, it is closely related to model ensembling [9], which averages over independently trained models rather than over transformations of a single input; both reduce variance in the final prediction.

Prior work establishes that TTA helps in aggregate pipelines, but rarely isolates how much it contributes on its own, and even more rarely measures whether that contribution is consistent across a controlled family of architectures and losses. The matched-condition models of [1] make such an isolated, controlled measurement possible, which is the gap addressed here.

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.v04i04.1471

3. Research Method

3.1 Models and Data

The sixteen trained models of [1] are reused without modification. Each pairs one of four U-Net-family decoders — U-Net, UNet++, MAnet, and LinkNet — with one of four segmentation losses (Dice, Focal, Tversky, and Focal-Tversky), all sharing a ResNet34 encoder pre-trained on ImageNet and identical preprocessing, training augmentation, and optimizer schedule. Evaluation is performed per slice on the held-out BraTS 2020 test set of 2,242 axial tumor-bearing slices. For two configurations (MAnet with Dice and MAnet with Focal-Tversky), batch-normalization running statistics are recomputed by a forward-only pass before evaluation, following [1], to correct a mixed-precision instability.

3.2 Test-Time Augmentation

For a model f and input slice x, the TTA prediction is the mean of the model's outputs over a set T of four label-preserving flip transforms — identity, horizontal flip, vertical flip, and both — each applied to the input and inverted on the output before averaging, as in Equation (1).

$$p_{TTA}(x) = \frac{1}{|T|} \sum_{t \in T} t^{-1}(f(t(x))) \quad (1)$$

Softmax probabilities are averaged across the four views at 32-bit precision, and the per-pixel class of maximum averaged probability is taken as the prediction. The no-TTA condition uses the identity transform only. The two conditions are otherwise identical, so any difference in Dice is attributable to the augmentation.

3.3 Evaluation and Statistical Testing

For each region r in {WT, TC, ET}, per-slice Dice is computed between predicted mask P and ground-truth mask G as in Equation (2), with the convention that a slice with no ground-truth and no predicted voxels of region r scores 1. The overall score for a slice is the mean of its WT, TC, and ET Dice.

$$\mathrm{Dice}_r = \frac{2\,|P_r \cap G_r|}{|P_r| + |G_r|} \quad (2)$$

For each model, the mean per-slice Dice with and without TTA is compared by a paired bootstrap over slices: for each of 10,000 resamples (fixed seed 42), slices are drawn with replacement and the mean per-slice Dice difference is computed. The observed difference Δ, its 95% confidence interval, and a two-tailed p-value are reported; a difference is significant when the interval excludes zero. A pooled analysis concatenates the per-slice differences of all sixteen models and applies the same bootstrap. As in [1], the slice-level bootstrap treats slices as independent and is therefore mildly anti-conservative, since slices from one patient are correlated.

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.v04i04.1471

4. Results

4.1 Every Model Improves Under TTA

Table 1 reports, for each of the sixteen models, the mean per-slice Dice without and with TTA, the difference, its 95% confidence interval, and its p-value. Every model improves, and in every case the 95% confidence interval excludes zero. Per-model gains range from +0.0054 (UNet++ with Dice) to +0.0220 (LinkNet with Focal). The pooled gain across all sixteen models is +0.0129 (95% CI [+0.0119, +0.0140], $p < 0.001$), raising pooled mean Dice from 0.813 to 0.826.

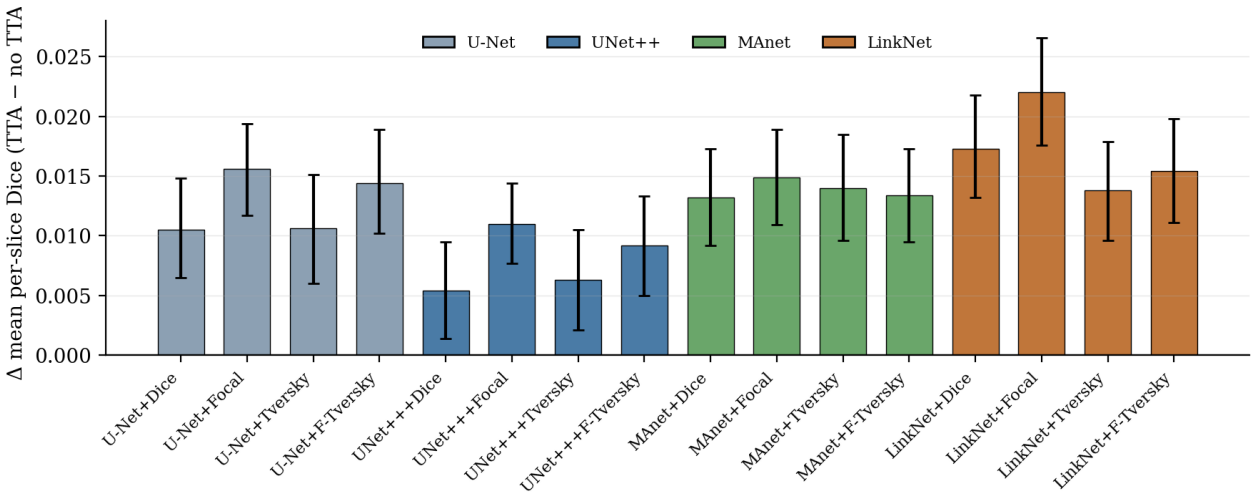
Table 1: Mean Per-Slice Dice Without and With Four-Flip Test-Time Augmentation for the Sixteen Matched Models, With the Paired-Bootstrap Difference, Its 95% Confidence Interval, and Two-Tailed p-Value. The Final Row Pools the Per-Slice Differences of All Sixteen Models.

Model	No TTA	TTA	Δ	95% CI	p
U-Net + Dice	0.8184	0.8289	+0.0105	[+0.0065, +0.0148]	< 0.001
U-Net + Focal	0.8108	0.8264	+0.0156	[+0.0117, +0.0194]	< 0.001
U-Net + Tversky	0.8106	0.8212	+0.0106	[+0.0060, +0.0151]	< 0.001
U-Net + Focal-Tversky	0.8152	0.8296	+0.0144	[+0.0102, +0.0189]	< 0.001
UNet++ + Dice	0.8258	0.8313	+0.0054	[+0.0014, +0.0095]	0.008
UNet++ + Focal	0.8240	0.8350	+0.0110	[+0.0077, +0.0144]	< 0.001
UNet++ + Tversky	0.8235	0.8298	+0.0063	[+0.0021, +0.0105]	0.004
UNet++ + Focal-Tversky	0.8087	0.8178	+0.0092	[+0.0050, +0.0133]	< 0.001
MAnet + Dice	0.8057	0.8189	+0.0132	[+0.0092, +0.0173]	< 0.001
MAnet + Focal	0.8111	0.8260	+0.0149	[+0.0109, +0.0189]	< 0.001
MAnet + Tversky	0.8141	0.8282	+0.0140	[+0.0096, +0.0185]	< 0.001

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.v04i04.1471

Model	No TTA	TTA	Δ	95% CI	p
MAnet + Focal-Tversky	0.8063	0.8196	+0.0134	[+0.0095, +0.0173]	< 0.001
LinkNet + Dice	0.8153	0.8327	+0.0173	[+0.0132, +0.0218]	< 0.001
LinkNet + Focal	0.8032	0.8252	+0.0220	[+0.0176, +0.0266]	< 0.001
LinkNet + Tversky	0.8056	0.8194	+0.0138	[+0.0096, +0.0179]	< 0.001
LinkNet + Focal-Tversky	0.8079	0.8233	+0.0154	[+0.0111, +0.0198]	< 0.001
Pooled (all 16)	0.8129	0.8258	+0.0129	[+0.0119, +0.0140]	< 0.001

Figure 1: Per-Model Improvement in Mean Per-Slice Dice From Four-Flip Test-Time Augmentation. Bars Show the Observed Difference Δ; Error Bars Show the 95% Bootstrap Confidence Interval. All Sixteen Bars Lie Entirely Above Zero. Colors Denote Decoder Family.



4.2 TTA Narrows the Gap Between Architectures

The improvement is not uniform across models. The largest gains occur for LinkNet (+0.0138 to +0.0220) and MAnet (+0.0132 to +0.0149), the decoder families that were weakest without augmentation, while the strongest family without TTA, UNet++, gains least (+0.0054 to +0.0110). Consequently, TTA compresses the spread of mean Dice across the sixteen models: configurations that trailed without augmentation are brought closer to the leaders. TTA therefore acts not only as a uniform accuracy boost but as a partial equalizer across decoder architectures.

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.v04i04.1471

4.3 Comparison With Model Ensembling

The pooled TTA gain of +0.0129 is of the same order as, and in fact exceeds, the ensemble gain of about +0.011 reported for a sixteen-model ensemble of these same configurations. The two techniques are related, as both improve predictions by averaging: TTA averages over input transformations of a single model, whereas ensembling averages over distinct models. TTA achieves its gain at the cost of a fixed multiple of a single model's inference and without storing or running multiple models, making it the more economical of the two prediction-averaging strategies.

5. Discussion

5.1 Interpreting the Result

The finding is a clean, consistent positive: a training-free inference-time technique improves every one of sixteen independently trained models, significantly in every case, with a pooled gain of about 1.3 Dice points. The consistency is the notable feature. Whereas varying architecture or loss in isolation did not reliably move aggregate accuracy in the matched comparison [1], TTA moves it in the same direction for every configuration. This is consistent with the variance-reduction view of prediction averaging: flip transforms produce correlated but non-identical predictions whose average is more stable than any single view.

5.2 Practical Implications

For practitioners, TTA is among the most favorable accuracy interventions available for this task: it requires no retraining, no architectural change, and no additional model storage, and it improved every model tested. Its cost is a fixed multiple of inference time (four forward passes for four flips), which is modest and parallelizable. Where inference budget is constrained, fewer transforms can be used; where a larger gain is sought, TTA can be combined with model ensembling, since the two operate on different axes of averaging.

5.3 Limitations

Three limitations bound these conclusions. First, evaluation and significance testing are performed at the slice level; because slices from one patient are correlated, the slice-level bootstrap is mildly anti-conservative, and patient-level aggregation would yield wider intervals, though given the size and unanimity of the effect the direction is not expected to change. Second, all experiments use a single dataset (BraTS 2020) and a 2D slice-based pipeline; generalization to 3D pipelines and other cohorts is untested. Third, only flip-based transforms are studied; richer augmentation sets (rotations, scaling, intensity transforms) may change the magnitude of the gain and are a natural extension.

6. Conclusion

Test-time augmentation was evaluated as an isolated, controlled intervention on sixteen matched U-Net-family configurations for brain tumor segmentation on BraTS 2020. Four-flip TTA improved mean per-

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.v04i04.1471

slice Dice for every model, with all sixteen improvements statistically significant and a pooled gain of +0.0129 ($p < 0.001$). The benefit was largest for the architectures that were weakest without augmentation, so TTA also narrowed the gap between decoder families. Because it requires no additional training and only a fixed multiple of inference cost, TTA is a reliable, economical, and broadly applicable improvement for U-Net-family brain tumor segmentation, and a natural complement to model ensembling.

7. Acknowledgement

The authors thank the organizers of the Multimodal Brain Tumor Segmentation (BraTS) challenge for making the BraTS 2020 dataset publicly available. This work reuses the trained models and evaluation infrastructure of the companion study [1]; no new model training was required. This work received no specific funding from any agency.

REFERENCES:

1. S. U. Ravi Kumar Chavali and P. V. Kumar, "A Controlled Comparison of U-Net Decoder Architectures and Loss Functions for Brain Tumor Segmentation on BraTS 2020," companion manuscript, 2024 (unpublished).
2. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, pp. 234–241, 2015.
3. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in Deep Learning in Medical Image Analysis (DLMIA), LNCS 11045, Springer, pp. 3–11, 2018.
4. T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation," IEEE Access, vol. 8, pp. 179656–179665, 2020.
5. A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation," in Proc. IEEE Visual Communications and Image Processing (VCIP), pp. 1–4, 2017.
6. 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.
7. F. Isensee, P. F. Jaeger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for Brain Tumor Segmentation," in Brainlesion (BraTS 2020), arXiv:2011.00848, 2020.
8. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks," Neurocomputing, vol. 338, pp. 34–45, 2019.
9. T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems, LNCS 1857, Springer, pp. 1–15, 2000.
10. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in Proc. 4th International Conference on 3D Vision (3DV), pp. 565–571, 2016.

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.v04i04.1471

11. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in Proc. IEEE ICCV, pp. 2980–2988, 2017.
12. S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," in Machine Learning in Medical Imaging (MLMI), LNCS 10541, Springer, pp. 379–387, 2017.
13. N. Abraham and N. M. Khan, "A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation," in Proc. IEEE International Symposium on Biomedical Imaging (ISBI), pp. 683–687, 2019.
14. B. H. Menze, A. Jakab, S. Bauer, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.