IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

A Comparative Study of a Hybrid CNN–Transformer Framework for Multimodal MRI-Based Brain Tumour Segmentation

Vikas Yadav and Ashwini Kumar Singh
Babu Banarsi Das University, Lucknow, Uttar Pradesh
vikasyadav.mca.ds@gmail.com
aashwiniisingh@bbdu.ac.in

Abstract: *Brain tumour segmentation in MRIs is a difficult task due to varying appearance of tumours between sequences, different regions in space and among cases. The CNNs are useful for efficient local feature extraction and decoding in multiple scales, whereas Transformer attention provides the ability to capture the relations between features in a wider sense. This work explores a compact U-Net architecture versus the hybrid CNN–Transformer model for a notebook on the basis of four modalities from the BraTS-GLI 2023 training collection: T1n, T1c, T2w and T2-FLAIR (T2f). Four-class pixel-wise segmentation of 128 × 128 resolution was performed by both models. Training was conducted using AdamW optimiser with weighted cross-entropy and foreground soft-Dice loss functions. Validation macro foreground Dice was used as a metric for checkpoints selection. Testing was conducted on 194 out-of-bag case IDs and six randomly chosen slices per case (total number of slices = 1,164 slices in total); therefore, the experiment is a selected-slice 2D evaluation rather than full-volume 3D segmentation.*

The two models scored a macro foreground Dice score of 0.7759 and 0.8132 on the chosen test slices, respectively, while their pooled whole-tumor Dice score was 0.8585 and 0.9139. The pixel accuracy scores of the U-Net and the hybrid were 0.9953 and 0.9969, respectively. A paired case-level comparison using 10,000 bootstrap samples, with the Wilcoxon signed-rank test and Holm correction showed significant differences that favoured the hybrid in dice scores of encoded label 2, macro foreground Dice, whole-tumor Dice, and pixel accuracy. However, there were no statistically significant differences between the two models for encoded labels 1 and 3 after Holm correction. The hybrid model contained 361,204 trainable parameters while those of the U-Net were 1,942,772. Although these results are promising, caution needs to be exercised since repeated numerically labelled data instances require identity verification, label semantics have to be validated against the official dataset documentation, no external dataset was used, and there is poor overlap of Grad-CAM with the reference foreground.

Keywords: Brain Tumour Segmentation, Multimodal Magnetic Resonance Imaging (MRI), BraTS-GLI 2023, Deep Learning, U-Net, Convolutional Neural Network (CNN), Transformer Architecture, Hybrid CNN–Transformer, Medical Image Analysis, Semantic Segmentation

I. INTRODUCTION

Brain tumor segmentation defines what parts of the brain are not normal based on medical images; it helps with quantitative studies, treatment planning and monitoring of disease progression over time. It is very time-consuming to manually segment the tumor area and the results are dependent upon how the person doing the segmenting interprets the image. Tumors tend to have ill-defined edges and appear differently in various types of images. Therefore,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

automated methods of segmenting tumors from medical images are critical areas of study in medical image analysis. However, achieving a high benchmark score in automated tumor segmentation does not mean that the algorithm will be effective in clinical practice, nor does it guarantee that the algorithm will be safe to use.

MRI has complementary contrasts. MRI BraTS-GLI 2023 data used in this paper comprises native T1-weighted (T1n), contrast enhanced T1-weighted (T1c), T2-weighted (T2w), and T2-FLAIR (T2f). The combination of these images enables the model to learn from different intensities and structures rather than depending on a single image. They are considered as four channels for input alignment in this experiment.

CNN encoder-decoder is still one of the good architectures for segmentation tasks. CNNs can effectively find local patterns, and skip connections will allow recovering spatial information during the decoding phase. The U-Net architecture introduced this kind of general architecture for biomedical image segmentation [1]. However, the CNN operation at each layer has a local receptive field, and in order to learn relations In case of far apart pixels, deep models are required. Self-attention allows representing features in a more extensive context, while attention blocks involve design trade-offs as well. Hybrid models try to combine the inductive bias of CNN with the context modelling of attention; however, this does not imply that all the hybrid models will always be superior to an optimised CNN model.

The current research focuses on comparing two specific models from the provided notebook – small U-Net and hybrid CNN-Transformer models with four channels of input, four output classes, four attention heads and two Transformer layers. Hybrid model is not considered a duplicate of TransUNet or Swin UNETR. The aim is to conduct performance comparison for the two models under the same data pre-processing, training procedure and slice selection tests.

A. Research aim and questions

The objective is to assess whether the developed hybrid CNN-Transformer model offers improved selected-slice segmentation accuracy in comparison with the developed U-Net baseline on the specified BraTS-GLI 2023 data split. The research questions are as follows: (1) What are the comparisons in encoded foreground class Dice, macro foreground Dice, pooled whole tumour Dice, and pixel accuracy? (2) Are there consistent case-wise differences through statistical paired testing corrected for multiple comparisons? (3) Is there compactness in terms of trainable parameter counts between the hybrid model and the U-Net? (4) Can exploratory Grad-CAM heatmaps be spatially aligned with the ground truth foreground?

B. Contributions

This paper presents an implementation-specific, controlled comparison using a frozen holdout test, case-level statistical analysis, reporting the number of parameters in models, and an exploratory analysis of Grad-CAM. This paper does not make any claims about presenting novel architecture or state-of-the-art performance, 3D segmentation, clinically validated results, or causality of the performance difference to Transformer in any way.

II. RELATED WORK

A. CNN encoder–decoder segmentation

U-Net is a neural network that uses the combination of a contracting path to learn context information and an expanding path to restore spatial information along with skip connections that propagate high-resolution information [1]. It represents one of the most popular starting points when developing networks for biomedical segmentation, which requires both local learning and spatial restoration. The current baseline is UNet2D notebook, which has a reported number of parameters and configuration, described in Section III.

B. Reproducible pipelines and nnU-Net

nnU-Net shows how the preprocessing, architecture selection, training, and post-processing steps can be adapted to fit a particular dataset and how much pipeline configuration influences segmentation quality [2]. This is important because architectural comparisons depend on preprocessing and training strategies. The current experiment is conducted using a

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

fixed 128 × 128 slice selection pipeline and specific architectures and does not involve nnU-Net or a comparison against an automatically configured 2D/3D pipeline from nnU-Net. Consequently, it provides an answer to a more limited question than a state-of-the-art benchmark.

C. Hybrid CNN–Transformer approaches

TransUNet utilizes CNN-based feature extraction along with Transformer-based encoding in order to incorporate long-range dependencies in medical image segmentation [3]. Swin UNETR leverages hierarchical shifted window Transformer representations within an U-Net structure and has been applied to brain tumor segmentation tasks [4]. Both papers provide motivation for the more general claim that attention mechanisms might supplement convolutional representations. However, they do not present direct architectural counterparts of the hybrid network in the notebook, as the proposed model relies on a significantly smaller number of parameters. Furthermore, comparisons to published results would be meaningless without identical training settings and benchmarks.

D. Multimodal MRI, dataset conventions and robustness

MRI Glioma segmentation usually makes use of more than one MRI sequence and analyses both the regions within the tumor and the union of these regions. The convention for releasing the dataset is important: the label information and the code corresponding to specific release should be verified instead of assuming it based on class IDs. BraTS 2023 dataset is available using the data infrastructure of the challenge [5]. Another paper dealing with glioma segmentation in realistic conditions under missing MRI sequences has shown that all modalities cannot be taken for granted in real life [6].

E. Explainability and Grad-CAM

The Grad-CAM technique applies the gradient flow into a chosen convolutional layer to create a coarse localization map [7]. Although it gives a qualitative representation of the spatial features affecting the target score, it is not a causal explanation and should not be confused with medical reasoning. In the present case study, the target class is picked using the ground-truth mask and its localization map is then thresholded and compared with the union foreground. This approach tests overlap under the circumstances when the target class is determined by the ground truth, not that the model independently recognizes a tumor.

F. Research gap addressed

Instead of making an assertion of the absence of any previous comparison, the current work is a study of reproducibility at a local level, asking the question of how two specific implementations of the model perform when applied to a certain selection of 2D slices after being trained and preprocessed using the same conditions. The limited nature of the experiment is pointed out in order to avoid confusion regarding benchmarking on full volumes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

III. MATERIALS AND METHODS

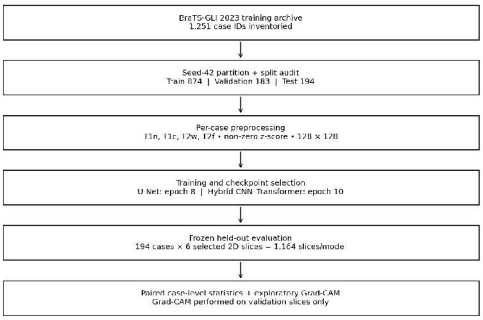


Fig. 1. Experimental workflow reconstructed from the notebook protocol. The Grad-CAM analysis uses validation slices, whereas aggregate segmentation metrics are computed on the frozen held-out test cache.

A. Dataset and provenance

The notebook indicates that the source archive is the BraTS-GLI 2023 training archive and the archive name is ASNR-MICCAI-BraTS2023-GLI-Challenge-TrainingData.zip. The supplied manifest indicates the presence of 1,251 unique case IDs. It is expected that each case would have four MRIs (t1n, t1c, t2w and t2f) and a segmentation volume. The terms of dataset access and the exact citation and license of the dataset should be confirmed from the upstream challenge documentation [5]. The data was not independently downloaded and re-audited for this manuscript preparation. The methodology and findings reported are taken from the supplied notebook.

B. Dataset inventory and split procedure

Partition	Cases	Share of 1,251	Role
Training	874	69.9%	Parameter optimisation
Validation	183	14.6%	Checkpoint selection and scheduler monitoring
Test	194	15.5%	Frozen final evaluation
Total	1,251	100%	Recorded case-ID inventory

The split procedure records random seed 42 and a grouped split manifest. The notebook reports zero overlap in numeric-base groupings across partitions, while also flagging 236 case records associated with 118 repeated numeric bases for identity verification. A numeric base is a grouping clue, not authoritative proof that records belong to the same or different individual. The current evidence therefore supports disjoint case identifiers and the recorded grouping check, but does not justify an unqualified claim of independent patient-level separation. This uncertainty is treated as a threat to validity.

The training/validation preprocessing manifest contains 1,057 cases, matching 874 training plus 183 validation cases. Test data were processed separately into a cache of 194 cases. The notebook's preflight and evaluation cells assert that test identifiers match the frozen test split and do not overlap the train/validation manifest by case ID.

C. Modalities and label encoding

Input data for the model is a four-channel tensor with channels T1n, T1c, T2w and T2f. Integer labels present in the segmentation mask include 0, 1, 2 and 3. In the primary table of results, labels 1-3 are preserved in the implementation and considered as foreground labels because the implementation itself does not specify their anatomical meaning. Labels 1, 2 and 3 are usually represented as necrotic/non-enhancing tumour core, peritumoral oedema and enhancing tumour correspondingly, while label 0 corresponds to the background. But this correspondence needs to be verified by the actual archive and its meta-data before assigning to the experiment [5]. Till then, this manuscript does not involve the anatomical relabelling. Whole tumour is the union of all non-zero labels.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

D. Preprocessing and normalisation

For the current case and modality, intensity mean and standard deviation are computed based on non-zero voxels only. In the case of non-zero voxel x, its normalised value will be $(x - \mu)/\sigma$, where μ and σ are the mean and standard deviation of non-zero voxels for the particular case and modality. Voxels that were equal to zero will retain their values. There is a special guard against near-zero standard deviations implemented in the notebook. It helps to reduce scale differences, while keeping zero-valued exterior as the zero background.

Four modalities are then normalised, stacked as channels and resized to 128 × 128 using bilinear interpolation with align_corners=False. Segmentation mask is resized to the same size using nearest-neighbour interpolation, while keeping integer class IDs instead of interpolating between them. There is no explicit augmentation pipeline in the inspected training code. The reduction of resolution from the original volume grid to 128 × 128 means also removal of spatial information, which could be an issue for small structures and boundaries. An example of the four modalities and encoded segmentation mask is shown in Figure 2.

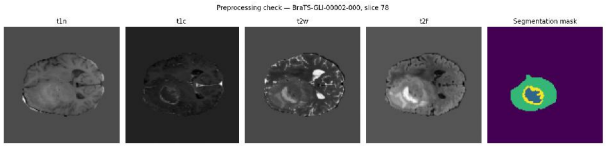


Fig. 2. Saved preprocessing example from the notebook: T1n, T1c, T2w and T2f channels with the corresponding encoded segmentation mask. This is a representative training case, not a test-set example.

E. Slice selection and test-cache construction

The experiment is conducted in 2D mode. In each instance, the code finds slices that contain at least one segmentation voxel other than zero and picks up four tumour-containing slices out of them, selecting them at approximately even distances from each other. Two slices without any tumours but still with some image content are chosen as well, applying the same even distance approach. Selected slices are made unique and sorted. There are six slices chosen in each instance, resulting in 194 × 6 = 1,164 slices in total in the test cache. The slice-selection process aims to cover both tumour-positive and non-tumour instances.

F. U-Net baseline

The starting point is the UNet2D model from the notebook. It takes four channels of inputs and four class predictions and has a base-width of 16. The number of trainable parameters reported is 1,942,772. No further inference is made on the layer-level details other than what is documented in the provided experiment summary.

G. Hybrid CNN–Transformer model

The second architecture is the HybridCNNTransformer for the particular notebook. Verified parameters include four input channels, four output channels, base number of channels set to 16, four Transformer attention heads, and two Transformer blocks. The reported number of trainable parameters is 361,204. The model performs a combination of convolution-based feature extraction and Transformer-based feature modeling to produce outputs of four-class pixel logits in the input image resolution. The provided information does not provide enough information about the particular architecture of the model in terms of the cells of model definition to perform layer-by-layer reconstruction, and it is advisable to keep the exact implementation along with the submitted code. This architecture is not TransUNet or Swin UNETR.

Property	U-Net baseline	Hybrid CNN–Transformer
Implementation class	UNet2D	HybridCNNTransformer
Input channels	4	4

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

Output classes	4	4
Base width setting	16	16
Attention heads	Not applicable	4
Transformer layers	Not applicable	2
Trainable parameters	1,942,772	361,204
Selected checkpoint epoch	8	10
Selection criterion	Validation macro foreground Dice	Validation macro foreground Dice

H. Training configuration and loss

Setting	**Recorded value**
Optimiser	AdamW
Initial learning rate	0.001
Weight decay	0.0001
Batch size	8
Requested training duration	10 epochs
Learning-rate scheduler	ReduceLROnPlateau
Scheduler mode / factor / patience	Maximise monitored metric / 0.5 / 2
Loss	Weighted cross-entropy + foreground soft-Dice loss
Class weights (0, 1, 2, 3)	[0.10, 1.50, 1.00, 1.50]
Gradient clipping	Maximum gradient norm 1.0
Random seed	42
Recorded GPU	NVIDIA Tesla T4
Recorded PyTorch	2.11.0+cu130

The cross entropy loss term uses the above-defined class weights. In the case of the Dice loss term, softmax outputs and one-hot targets are averaged across the batch and spatial dimensions. Foreground Dice loss is calculated for the foreground classes available in the ground truth batch, while the background is excluded. The total loss is the weighted cross entropy plus one minus the average Dice score for the present foreground classes. Gradients' L2 norm is clamped to 1.0 prior to the optimiser update. The learning rate for the hybrid model as stored in the training log is 0.001 through all the requested ten epochs.

I. Checkpoint selection and test isolation

For each model, the checkpoint with the highest validation macro foreground Dice is selected. The selected U-Net checkpoint is for epoch 8 and the selected hybrid checkpoint is for epoch 10. Both checkpoints have the same seed 42 stored in the notebook. The checkpoints are frozen prior to evaluation on the held-out test set. The test set has not been used in checkpoint selection or model tuning; this is an advantage even though a single seed does not cover variability from different initialisation and data orderings. A prediction sample forample is shown in Fig. 3; it is illustrative and is not a held-out test case.

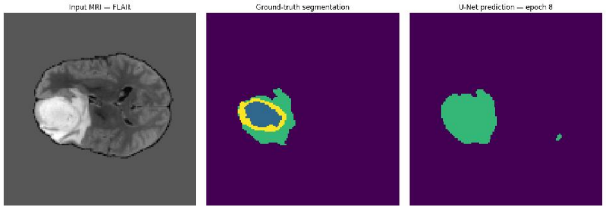


Fig. 3. Saved U-Net validation example from the notebook (epoch-8 checkpoint): FLAIR input, ground-truth mask and predicted mask. This qualitative example is validation-only and should not be interpreted as a held-out test result.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

J. Evaluation metrics

In terms of a class-based binary comparison, TP is true-positive pixels, FP is false-positive pixels and FN is false-negative pixels. The Dice coefficient is:

$$\text{Dice} = 2TP / (2TP + FP + FN). \quad (1)$$

Precision and recall are defined as follows:

$$\text{Precision} = TP / (TP + FP). \quad (2) \qquad \text{Recall} = TP / (TP + FN). \quad (3)$$

The Intersection over Union (IoU) metric is defined as:

$$\text{IoU} = TP / (TP + FP + FN). \quad (4)$$

Pixel accuracy is the ratio of all correctly classified pixels among the total number of evaluated pixels:Pixel accuracy = (number of correct pixels) / (total number of pixels). (5)Macro foreground Dice is the arithmetic mean of Dice scores for encoded labels 1, 2 and 3 accounting for finite-value calculations in the implementation. Whole-tumour Dice is determined by pooling all pixels belonging to the union of non-zero labels: predicted foreground = prediction > 0 and true foreground = target > 0. The aggregated class Dice and whole-tumour Dice reported are pooled across the selected test pixels, and not the mean of case-level metrics. In contrast to this, paired inference is based on only one metric value per case, which is compared between the hybrid and U-Net results. These two aggregation approaches address distinct problems and cannot be mixed.

In the case where neither prediction nor ground truth contains any pixels from a certain class, class Dice is undefined and reported as missing, not equal to one. As a result, paired sample sizes for labels 1 and 3 are less than 194. Pixel accuracy is additionally reported as a secondary metric, since the majority of pixels belong to the background and may dominate the numerator even in case of inaccurate foreground borders.

K. Statistical analysis

For each metric, the paired difference is $\Delta_i = \text{metric}_i,\text{hybrid} - \text{metric}_i,\text{U-Net}$ for each case with valid values from both models. The mean and median paired differences were calculated. A non-parametric bootstrap with 10,000 resamples and random seed 42 estimated a 95% percentile confidence interval for the mean paired difference. Two-sided Wilcoxon signed-rank tests were used for paired comparisons, and Holm correction was applied across six metrics. P-values are reported as $p < 0.001$ where the computed output rounded to zero; they are not interpreted as exactly zero. Statistical significance is not equivalent to clinical importance.

L. Grad-CAM analysis

Grad-CAM Exploration was performed on the hybrid model with target layer being enc3.block.3. Target class for each valid validation slice was determined based on the largest foreground class in the corresponding ground-truth mask. Target score was calculated as the mean of the target-class logit over the foreground pixels in the ground-truth mask. Channel weights were calculated by spatially averaging gradients and the result of the weighted activations sum after applying ReLU activation was bilinearly upscaled to 128 x 128 size and normalised to the range [0, 1]. Quantitative subset of images included the first 100 valid validation slices containing at least 50 foreground pixels. Activation maps were thresholded and compared with the union foreground mask (ground-truth > 0). As the target selection is based on ground-truth, this is not a tumour-discovery test.

M. Hardware, software and reproducibility

This notebook contains an NVIDIA Tesla T4 GPU along with PyTorch 2.11.0+cu130. NumPy, pandas, SciPy, nibabel, and other libraries are used in this notebook; however, pinned version numbers have not been set up for all of the dependencies in the current record. The seed is set to 42 for Python/NumPy/PyTorch in training functions. Manifests, caches, checkpoints, and evaluation results are stored in a runtime environment of the notebook located in the /content folder. Reproducible execution necessitates that the notebook itself, the version of the source archive, the split manifest, preprocessing manifests, model checkpoints, packages version numbers, and output files be kept.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

N. Data-leakage safeguards

Workflow

- Separate train/validation preprocess from test preprocess.
- The case-ID splits should have no overlaps.
- Test cache IDs must match the frozen test split.
- Frozen checkpoints are evaluated without any test-based selection.

These precautions prevent the leakage of case-IDs directly. They fail to determine whether the reuse of numerical base IDs indicates common patient identity. They also fail to prove independence when the official patient ID is used.

The issue remains unresolved until verification of metadata is done..

IV. EXPERIMENTAL RESULTS

A. Evaluation inventory

The final test evaluation comprised 194 case IDs and six selected slices per case for each model, amounting to 1,164 slices per model. With slices at 128 x 128 pixels and 4-class masks, the notebook indicates that the evaluation includes 19,070,976 evaluated pixels in total across the selected slices. The evaluation was conducted on both frozen checkpoints.

B. Aggregate metric comparison

Metric	U-Net	Hybrid CNN–Transformer
Dice, encoded label 1	0.7169	0.7510
Dice, encoded label 2	0.7671	0.8289
Dice, encoded label 3	0.8438	0.8596
Macro foreground Dice	0.7759	0.8132
Pooled whole-tumour Dice	0.8585	0.9139
Pixel accuracy	0.9953	0.9969

Higher overall results can be observed for the hybrid model in Table 3 on all six metrics. There was an increase in macro foreground dice of 0.0373, as well as an increase in pooled whole-tumour dice of 0.0553. This represents the difference that is observed in a descriptive manner based on the selected-slice approach. Pooled whole-tumour score represents the aggregation of foreground intersection and denominator over all pixels tested; it is not the average per-case whole-tumour Dice score. High pixel accuracy scores need to be considered with Dice because of a lot of correctly classified background pixels may overshadow the foreground classification accuracy.

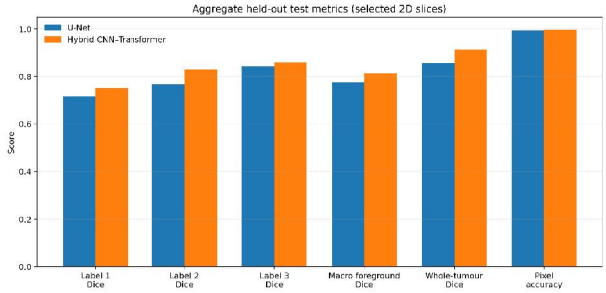


Fig. 4. Aggregate test metrics for the U-Net and hybrid CNN–Transformer on 1,164 selected 2D slices per model. Pixel accuracy is shown on the same 0–1 scale and should be interpreted cautiously because of background dominance.

C. Class-wise Dice

The pooled Dice score of label 1 is 0.7169 for U-Net and 0.7510 for the hybrid model. The pooled Dice scores for label 2 are 0.7671 for U-Net and 0.8289 for the hybrid. The pooled Dice scores for label 3 are 0.8438 and 0.8596

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

respectively for U-Net and the hybrid. The maximum absolute difference of the class scores for the hybrid is observed for label 2. These labels are deliberately not mapped to any anatomical name since the exact convention of the dataset labels needs verification.

D. Paired statistical results

Metric	n	Mean Δ	Median Δ	Bootstrap 95% CI for mean Δ	Holm-adjusted p
Encoded label 1 Dice	187	0.016041	0.000000	[−0.006160, 0.039327]	0.435935
Encoded label 2 Dice	194	0.054696	0.035523	[0.041985, 0.067519]	< 0.001
Encoded label 3 Dice	192	0.010636	0.000000	[−0.002183, 0.024291]	0.147349
Macro foreground Dice	194	0.036280	0.011715	[0.022589, 0.051219]	0.000001
Whole-tumour Dice	194	0.061519	0.032987	[0.050025, 0.074799]	< 0.001
Pixel accuracy	194	0.001588	0.000936	[0.001288, 0.001904]	< 0.001

There was a statistical significance in Dice scores of encoded label 2, foreground macro Dice score, whole-tumour Dice score, and pixel accuracy after Holm correction. The confidence intervals for mean difference between the metrics were above zero. However, confidence intervals for mean difference between metrics of encoded labels 1 and 3 contained zero, and adjusted p-values were not significant. Therefore, the findings do not indicate that each foreground class has improved significantly. Statistical significance in change in pixel accuracy does not imply any improvement in segmentation quality, especially under the condition of class imbalance.

The sample sizes in class-specific paired comparisons are different since class Dice scores at the level of cases are not defined in cases where the respective class is absent in the ground truth and predicted images. Thus, the sample sizes are 187 for label 1 and 192 for label 3

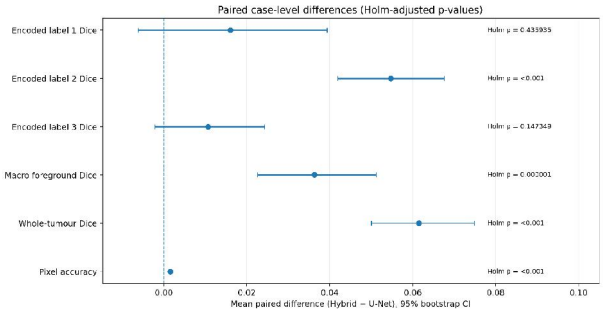


Fig. 5. Mean paired differences (hybrid minus U-Net) with 95% bootstrap confidence intervals. Intervals crossing zero align with the non-significant corrected results for encoded labels 1 and 3.

E. Direction of case-level differences

Metric	Hybrid better	Tie	U-Net better	Paired cases
Encoded label 1 Dice	79	23	85	187
Encoded label 2 Dice	164	0	30	194
Encoded label 3 Dice	95	7	90	192
Macro foreground Dice	128	0	66	194
Whole-tumour Dice	182	0	12	194
Pixel accuracy	176	0	18	194

While the direction count is supplementary to the confidence interval and the corrected test of hypotheses, the latter two are still preferred. The hybrid algorithm performed better at the tumour level for 182 out of 194 cases and at the macro foreground Dice score for 128 out of 194 cases. For encoded class 1, there were more favourable cases for U-Net than

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

the hybrid algorithm (85 cases versus 79 cases), with 23 ties, and no statistically significant mean difference. Encoded class 3 was also balanced with regard to the direction count and not significantly adjusted.

F. Grad-CAM threshold sensitivity

CAM threshold	Mean precision	Mean recall	Mean Dice	Mean IoU	Median Dice
0.25	0.1221	0.1109	0.0968	0.0668	0.0000
0.50	0.1326	0.0574	0.0708	0.0461	0.0000
0.75	0.1413	0.0234	0.0379	0.0222	0.0000

The precision value was higher with a more stringent threshold, while the recall, Dice and IoU values were lower. Median values were zero for all the three thresholds considered. This suggests poor spatial overlap between the CAM activation map and the ground truth foreground using the validation sample. However, the findings cannot imply that Grad-CAM is not useful in all contexts but cannot prove the validity of Grad-CAM in terms of tumour localization or explanations that have clinical validity. The findings are dependent on the target layer, the target class using ground truth information, score calculation, normalization, and thresholding.

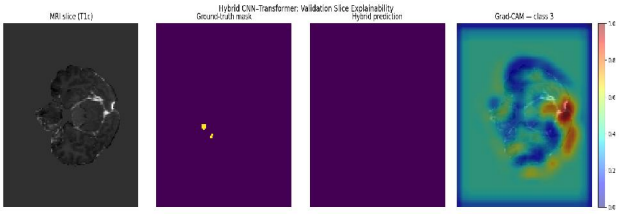


Fig. 6. Saved single-slice Grad-CAM output for the hybrid model on a validation slice. The explained foreground class was selected using the ground-truth mask; the heatmap is exploratory and is not a clinically validated explanation.

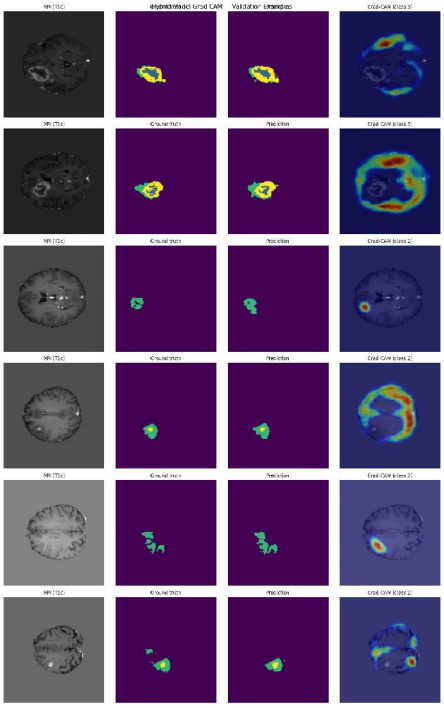


Fig. 7. Saved multi-slice Grad-CAM examples from validation data. The examples illustrate variability in saliency maps and are not evidence that Grad-CAM independently discovers tumour regions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

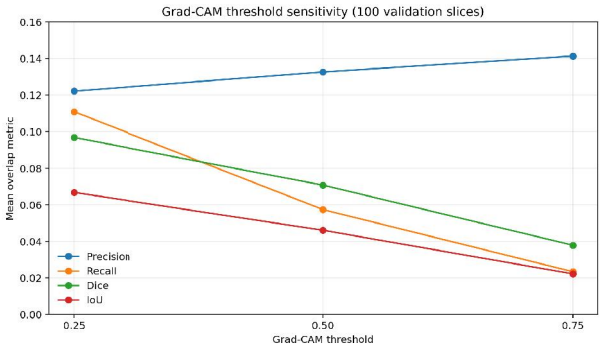


Fig. 8. Mean thresholded Grad-CAM overlap metrics on 100 eligible validation slices. All median metrics were zero at each threshold, indicating weak and sparse spatial overlap under this evaluation procedure.

V. DISCUSSION

The hybrid approach outperformed the U-Net in all published metrics of aggregated statistics and achieved paired improvements in encoded label 2 Dice score, macro foreground Dice, whole tumour Dice, and pixel accuracy with statistical significance. Positive difference for the whole-tumour case level metric occurred in 182 of 194 cases, which is in line with the positive interval for the bootstrap test and corrected Wilcoxon test results. According to this protocol, there is enough evidence to conclude that the hybrid approach outperformed the U-Net on aggregate foreground task and encoded label 2 comparison.

The improvements were not uniform. Paired differences in encoded label 1 and label 3 were insignificant after Holm correction and had confidence intervals containing zero. These two facts were supported by the direction counts: the U-Net outperformed the hybrid approach on the encoded label 1 in more cases, and label 3 comparisons were nearly balanced. It should be noted that macro scores could be improved even if each of the classes did not show improvements because the macro metric was a sum of performance of all classes and whole tumour metric was a merger of all non-zero classes into one foreground region.

The parameter count of the hybrid model is 361,204, which is about 18.6% of the parameter count of the U-Net (1,942,772). This indicates that the hybrid model may require fewer parameters for this particular implementation. However, this cannot be used to claim that the hybrid model has lower latency, lower peak memory, lower energy consumption, or faster inference speed because these metrics depend on feature map sizes, attention computations, and hardware implementations, and hence were not calculated in this experiment.

It must be noted that the performance increase cannot be attributed solely to the use of Transformer attention because this is a comparison of two complete model architectures, and thus they may have differences in features pathways and architecture. To prove this point, an ablation study is necessary. Furthermore, these models were trained for ten epochs and only one random seed was used. There is no external validation and hence the ranking can change with further training or using different seeds.

For the hybrid approach, pixel accuracy reached 0.997; however, this measureis biased towards the background class in a segmentation problem where majority of the pixels are not tumour. Dice and overlap measures for region-level segmentation are informative when applied to foreground segmentation, but they are also dependent on the presence of the class of interest and how aggregation is performed.

The Grad-CAM analysis is exploratory and lacks quantitative power. Since the target class is defined using the ground-truth mask, the method conditions the saliency generation process on knowledge about the target region. Low average Dice and zero median Dice scores for every threshold value indicate that there is no overlap between the foreground reference and saliency map when a simple thresholding is used. Saliency maps can be rough and sensitive to the choice

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

of layers, gradients, and thresholds; they cannot be used as proof of clinical reasoning ability, safety, and independent lesion detection.

The above discussion shows that there is enough proof regarding the implementation-specific performance difference of a selected 2D slice benchmark. The clinical significance, generalisation across other hospitals and scanner models, and volume segmentation will have to be demonstrated in future studies..

VI. LIMITATIONS AND THREATS TO VALIDITY

First, only six selected 2D slices per case were evaluated. This is not exhaustive inference across the 3D volume and may over-represent selected tumour-positive and selected background slices. The scores cannot be directly interpreted as full-volume segmentation performance.

Second, repeated numeric-base identifiers remain unresolved. Although the split manifest reports no numeric-base overlap, 236 case records associated with 118 repeated numeric bases require authoritative verification. Until the official metadata clarifies the identity relationship, patient-level independence should not be claimed.

Third, the encoded label semantics must be verified against the official BraTS-GLI 2023 release documentation. The current results are labelled 1–3 and avoid anatomical claims. The official dataset convention may offer a likely mapping, but the exact archive version and any preprocessing changes should be checked before final manuscript submission.

Fourth, the training was limited to ten epochs of request and one random seed number. It cannot quantify any stochasticity in the training nor provide proof of convergence. Fifth, an external validation dataset was not used, thus making generalization to other datasets unknown.

Sixth, all the four modalities were present in the experimental setup. There was no such experiment on missing modality or corrupted inputs. Seventh, there was no defined augmentation pipeline present in the notebook.Eighth, Grad-CAM was evaluated on a subset of 100 eligible validation slices and showed weak thresholded overlap. The target class was selected using ground-truth information, and the results are not clinical validation of explanation quality. Ninth, class imbalance can inflate pixel accuracy and should be considered when interpreting the high accuracy values.

Tenth, statistical significance does not mean clinical significance. The study performs a test on six parameters with Holm's correction but no threshold value for clinical relevance is given. Eleventh, the comparison is between only two versions with no more baselines including nnU-Net, TransUNet and Swin UNETR that can be considered a state-of-the-art comparison.

VII. CONCLUSION AND FUTURE WORK

This study conducted a comparison between a compact 2D U-Net and a notebook-oriented hybrid CNN-Transformer on four-modality MRI segmentation using 194 cases and six slices per each case. Hybrid network performed better on all three foreground label Dice metrics, macro foreground Dice, pooled whole-tumor Dice and pixel accuracy metrics. Pairs-wise comparison with bootstrapped confidence intervals, Wilcoxon tests and Holm correction found significant differences for Dice metric of encoded label 2, macro foreground Dice, whole tumor Dice and pixel accuracy, but not for Dice metrics of encoded labels 1 and 3. Hybrid network also had less trainable parameters than U-Net.

These results are valid only for the described slice selection procedure and cannot be regarded as evidence of 3D performance, clinical relevance or overall superiority of Transformer-based segmentation method. Repeated-ID identity relation and official label mapping have to be verified. The spatial localisation task was not successful in the experiment with Grad-CAM.

Future work needs to validate full 3D volumes, validate identity and labeling of the cases from official metadata, training over several seeds, and validation on an independent dataset. The robustness check needs to account for missing MRI modalities and intensity variability. The computational comparison needs to report latency, maximum resource utilization under controlled hardware conditions. The explainability evaluation needs to adopt a stricter

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

procedure that involves sensitivity analysis and localization metrics. Lastly, more baselines need to be benchmarked with nnU-Net and hybrid models.

ACKNOWLEDGEMENT

In completing the research work titled "A Comparative Study of a Hybrid CNN-Transformer Framework for Multimodal MRI-Based Brain Tumour Segmentation," I have had the guidance, encouragement, and support of many people. I feel privileged to thank all those people who have in one way or another contributed in making this research work a success.

First of all, I would like to thank my research supervisor, Mr. Rakesh Ranjan, for his valuable guidance, suggestions, constructive criticism, and encouragement throughout this research work. The insights, suggestions, and guidance of my research supervisor were very important in understanding deep learning and brain tumour segmentation. The guidance of Mr. Rakesh Ranjan was very helpful in conducting the experiments and completing this research.

Next, I would like to thank Prof. Dr. Reena Srivastava, Hon'ble Dean, for her valuable encouragement and guidance throughout my academic life.

Moreover, I am grateful to Prof. Dr. Manish Kumar, Head of the Department, for his academic guidance and valuable suggestions.

I would also like to express my appreciation to all the faculty members of the School of Computer Applications, Babu Banarasi Das University, Lucknow, for their knowledge, encouragement, and academic support throughout my studies.

I am grateful to the researchers and institutions responsible for developing and maintaining the BraTS-GLI 2023 dataset, which provided the multimodal MRI data used in this study. I also acknowledge the developers and contributors of the open-source libraries and deep learning frameworks that supported the implementation, training, evaluation, and visualisation of the U-Net and Hybrid CNN–Transformer models.

Finally, I express my heartfelt gratitude to Almighty God, my parents, family members, and friends for their unconditional love, motivation, patience, and unwavering support. Their encouragement has been a I also want to thank all the faculties of the School of Computer Applications, Babu Banarasi Das University, Lucknow, for their expertise, motivation, and academic support during my studies.

I am thankful to the organizations behind the development and maintenance of the BraTS-GLI 2023 dataset, which made it possible to use multimodal MRI datasets in this research work. I am also thankful to the developers of various open-source libraries and deep learning frameworks, which helped to implement, train, evaluate and visualize the results of the U-net and Hybrid CNN-transformer models.

Lastly, I am thankful from the bottom of my heart to Almighty God, my parents, family members, and friends for their unconditional love, motivation, patience, and constant support. Their motivation has been my constant source of energy throughout my academic journey.

Vikas Yadav
MCA (Data Science & Artificial Intelligence)
School of Computer Applications
Babu Banarasi Das University
Lucknow, Uttar Pradesh, India.

DECLARATION

I, Vikas Yadav, MCA Data Science and Artificial Intelligence student in the School of Computer Applications of Babu Banarasi Das University, Lucknow, Uttar Pradesh, hereby certify that the research entitled "A Comparative Study of a Hybrid CNN–Transformer Framework for Multimodal MRI-Based Brain Tumour Segmentation" has been performed under the guidance of Mr. Rakesh Ranjan.

In this research, a comparative study has been performed on U-Net and Hybrid CNN–Transformer models for brain tumour segmentation using multimodal MRI scans obtained from the BraTS-GLI 2023 dataset. The process of research

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

is involved in data preparation, model implementation, experiments, comparing the performances of segmentation and analysing the results.

I am taking the full responsibility for the correctness of experimental results, their analysis and interpretation, as well as for correct citation of all sources of information used in my research. All necessary publications, datasets, software packages and other external sources used in this research have been correctly cited in accordance with academic traditions.

Furthermore, I declare that the paper will be submitted in strict compliance with the academic integrity, originality, authorship and AI-assisted writing policies of Babu Banarasi Das University and the intended publication venue.

Author: Vikas Yadav
Programme: MCA (Data Science & Artificial Intelligence)
Research Supervisor: Mr. Rakesh Ranjan
Department: School of Computer Applications
Institution: Babu Banarasi Das University
Location: Lucknow, Uttar Pradesh, India

DATA AVAILABILITY

In this research, the BraTS-GLI 2023 MRI brain tumour dataset is employed in developing and experimenting with the models. The dataset has been obtained from the sources available via the dataset providers and used in compliance with the policies on accessing the dataset. Further access to the dataset and usage thereof is conditioned upon the requirements set forth by the dataset providers.

Dataset Repository: https://www.synapse.org/Synapse:syn51514105
Challenge Information: https://www.synapse.org/Synapse:syn51156910

ETHICAL CONSIDERATIONS

In this study, deep learning models have been used to analyze an existing MRI dataset pertaining to brain tumours. This research does not entail any experimentation, clinical treatment, or any involvement of human participants by the researcher. The MRI dataset was used for the purpose of conducting research according to the dataset conditions that apply.

The researcher appreciates the need to comply with the research ethics policy of the institution and the permissions given for the use of the dataset. It is important to verify any other approval or exemption necessary for the submission to the university and for publication at a journal.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, LNCS, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[2] 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, doi: 10.1038/s41592-020-01008-z.

[3] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv:2102.04306, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306. Bibliographic version and publication details should be checked before submission.

[4] A. Hatamizadeh et al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," arXiv:2201.01266, 2022. [Online]. Available: https://arxiv.org/abs/2201.01266. Verify the final citation version required by the university.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

[5] ASNR–MICCAI, "BraTS 2023 Adult Glioma (GLI) Challenge: Data and Terms." Official challenge documentation and data-access pages. [Online]. Available: https://www.synapse.org/Synapse:syn51156910. Confirm the exact dataset citation, version, licence and label convention against the accessed archive before submission.
[6] R. Pemberton et al., "Multi-class glioma segmentation on real-world data with missing MRI sequences: Comparison of three deep learning algorithms," Scientific Reports, vol. 13, Art. no. 18911, 2023, doi: 10.1038/s41598-023-44794-0.
[7] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[8] S. Bakas et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," Scientific Data, vol. 4, Art. no. 170117, 2017, doi: 10.1038/sdata.2017.117