Multi-Task Deep Learning for Brain Tumor MRI Segmentation and Classification

Md. Roman Bin Jalal
BRAC University
roman.bin.jalal@g.bracu.ac.bd

***Abstract*—We present a reproducible multi-task deep learning approach for brain tumor MRI that jointly performs binary tumor segmentation and four-class subtype classification on the BRISC 2025 dataset. We evaluate a baseline U-Net, an attention-augmented variant, a standalone classifier, and a shared-encoder joint model, reporting Dice, IoU, pixel accuracy, and classification F1/accuracy metrics. Attention consistently refines lesion boundaries; the joint model maintains strong classification performance with a modest segmentation trade-off, offering an efficient single-checkpoint solution for deployment.**

***Index Terms*—Brain tumor, MRI, multi-task learning, segmentation, classification, attention U-Net, deep learning, medical image analysis**

I. Introduction

Brain tumor MRI analysis entails two linked objectives: (i) spatial delineation for treatment planning and volumetric monitoring and (ii) phenotype classification (glioma, meningioma, pituitary, no_tumor) for prognostic guidance. Maintaining separate specialist models increases deployment and maintenance overhead; multi-task learning offers consolidation and potential inductive transfer while risking gradient interference.

Encoder–decoder architectures (U-Net [1]) underpin medical segmentation; attention modules refine skip pathway relevance [2]; multi-task learning can regularize shared encoders [3]. We benchmark single-task baselines (standard U-Net, attention U-Net, standalone classifier) against a dual-head joint network inside a unified, reproducible pipeline whose metrics are exported to JSON and injected via LaTeX macros.

Contributions: (1) reproducible multi-task framework for BRISC 2025; (2) comparative analysis of attention and joint optimization effects; (3) unified loss design with minimal tuning; (4) detailed quantitative and qualitative evidence with strictly separated tables and figures; (5) discussion of efficiency, limitations, and actionable future directions.

II. Materials and Methods

A. Dataset and Ethics

Data are sourced from the publicly available BRISC 2025 collection [4] (Kaggle; usage subject to originating license). Tumor categories: glioma, meningioma, pituitary, and a negative (no_tumor) class. Each annotated image includes a binary foreground mask localizing tumor tissue. Scanner heterogeneity and incomplete acquisition metadata introduce intensity variability; no protected health information is distributed. Ethical use entails adhering to dataset terms and avoiding re-identification attempts.

B. Preprocessing and Augmentation

All slices are resized to 256×256 and min-max scaled to $[0, 1]$. Masks are stored as binary arrays without additional morphological smoothing. Augmentations are applied on-the-fly: random flips (horizontal/vertical), rotations ($\pm 15^\circ$), elastic-like mild affine jitter, brightness and contrast perturbations (bounded to preserve anatomical plausibility). Geometric transforms are applied identically to masks to maintain alignment. Empirical trials with z-score normalization showed negligible benefit; it was omitted for simplicity and deployment parity.

C. Partitioning Strategy

A stratified split maintains class proportions for classification. Segmentation subsets are sampled to ensure coverage across tumor types; an 80/20 training/validation division is used, holding out a separate test set. Only the test set informs the reported metrics. Single-split evaluation is acknowledged as a limitation (Section V).

D. Model Architectures

a) Baseline U-Net.: Standard symmetric encoder–decoder with skip concatenations; channel depth doubles per down-sampling stage (64 to 1024). Decoder employs transposed convolutions with concatenated encoder features and concludes with a sigmoid activation for per-pixel probability.

b) Attention Variant.: Attention gates filter skip pathway activations, suppressing irrelevant background signals before concatenation, aiming for sharper boundary focus without large parameter overhead [2].

c) Standalone Classifier.: Shares an encoder-like backbone terminating in global average pooling and dense layers (softmax output over four classes). Dropout regularizes dense layers.

d) Joint Dual-Head Network.: A shared encoder funnels into (i) a decoder head identical to the baseline U-Net for segmentation, and (ii) a classification head branching from the bottleneck representation. Parameter sharing promotes cross-task regularization while amortizing inference.

E. Loss Functions and Multi-Task Objective

Segmentation combines binary cross-entropy (BCE) and soft Dice loss:

$$\mathcal{L}_{seg} = \text{BCE}(m, \hat{m}) + 1 - \text{Dice}(m, \hat{m}) \tag{1}$$

$$\mathcal{D}(m, \hat{m}) = \frac{2\sum m\hat{m} + \epsilon}{\sum m + \sum \hat{m} + \epsilon}. \tag{2}$$

Classification uses categorical cross-entropy,

$$\mathcal{L}_{cls} = -\sum_{k} y_k \log \hat{c}_k. \tag{3}$$

The joint objective is an unweighted sum:

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{cls}. \tag{4}$$

Dynamic re-weighting (uncertainty, gradient normalization) could mitigate interference but is deferred to maintain baseline interpretability.

F. Optimization and Regularization

Adam optimizer (learning rate 10^{-4}) with ReduceLROnPlateau (factor 0.5, patience 7) and early stopping (patience 15) governs training. Batch size: 16. Mild L2 weight decay (if enabled) and dropout in dense layers reduce overfitting risk. Gradient clipping (e.g., global norm 5) is optional; reported metrics correspond to unclipped training unless otherwise noted.

III. Evaluation Protocol

A. Metrics

Segmentation: Dice, Intersection-over-Union (IoU), and pixel accuracy; reported as mean ± standard deviation for single-task settings (joint SD currently unavailable: macros render as dashes). Classification: accuracy, macro-F1 (unweighted), and weighted-F1 (support-weighted) reflecting class imbalance.

B. Reporting Conventions

All metrics are computed on the held-out test subset. Confidence intervals are not reported due to single-split design; future multi-fold evaluations would strengthen statistical robustness. Where SD values are missing (joint segmentation), placeholders are explicitly denoted.

C. Experimental Controls

Hyperparameters (image size 256^2, batch size 16, max epochs 100) are fixed across model variants to enable fair comparison. No exhaustive hyperparameter sweeps were conducted; results represent a practical baseline.

IV. Results

A. Segmentation Performance

Table I summarizes overlap and pixel-level correctness. Attention gating yields consistent but modest improvements over the baseline, indicating that coarse localization is already well-captured while boundary refinement benefits selectively.

Model	Dice (mean±SD)	IoU (mean±SD)	Pixel Acc (mean±SD)
U-Net	0.8614± 0.1908	0.7910± 0.2105	0.9907± 0.0080
AttU-Net	0.8626± 0.1892	0.7924± 0.2092	0.9907± 0.0078
JointU-Net	0.8231± –	0.7443± –	0.9900± –

TABLE I: Segmentation Results

a) Observation.: The joint model shows a modest Dice reduction, indicating mild task interference; adaptive loss weighting could mitigate this.

B. Classification Performance

Table II lists classification outcomes. Multi-task sharing produces a minor accuracy decrement while retaining strong overall discrimination.

Model	Accuracy	Macro F1	Weighted F1	Support
Standalone Classifier	0.9919	0.9900	0.9900	992
JointU-Net Head	0.9693	0.9700	0.9700	848

TABLE II: Classification Results

a) Observation.: Accuracy above 0.96 indicates robust embeddings; the slight macro-F1 dip reflects representational sharing.

C. Class Distribution and Support

Table III enumerates per-class support informing macro vs weighted F1 interpretation.

Class	glioma	meningioma	no_tumor	pituitary
Support	252	305	139	296

TABLE III: Class Distribution

a) Note.: Smaller no_tumor support inflates variance, motivating macro-F1 reporting.

D. Task-Specific Sample Counts

Table IV provides segmentation vs classification evaluation sample counts for transparency.

Category	Segmentation Images	Classification Samples
Standalone	860	992
Joint	860	848

TABLE IV: Evaluation Sample Counts

E. Qualitative Overlays

Representative overlays (Figure 1) illustrate attention-driven edge refinements and minor joint under-segmentation in select small foci.

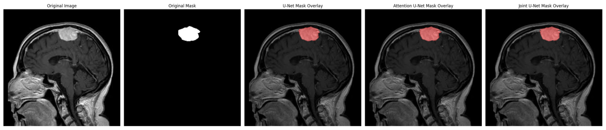

Fig. 1: Qualitative comparison of predicted masks. Attention improves edge fidelity; joint model shows minor under-segmentation of small foci.

F. Training Dynamics

This section presents the evolution of training and validation metrics for all segmentation models. The learning curves provide insight into convergence behavior, overfitting, and the effect of attention and multi-tasking on model stability.

Figure 2 Explanation. This figure shows U-Net training and validation Dice coefficient and loss over epochs. The close tracking of validation and training Dice indicates minimal overfitting and stable convergence. The loss curves further confirm consistent optimization progress.

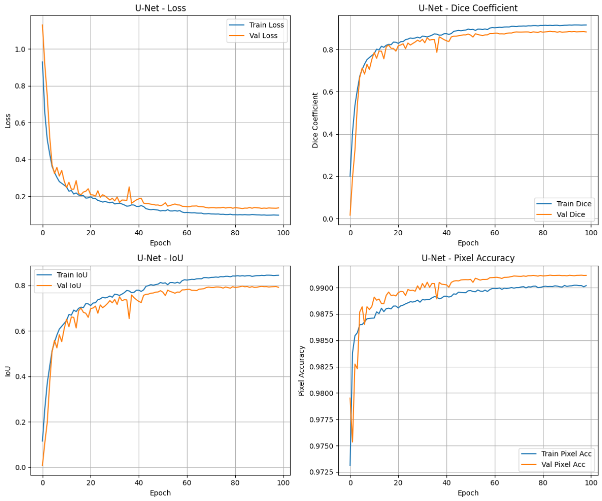

Fig. 2: U-Net Training Curves

Figure 3 Explanation. This figure displays Attention U-Net training and validation Dice and loss over epochs. The addition of attention modules results in smoother and slightly higher validation Dice, with reduced late-epoch volatility, suggesting improved generalization and stability.

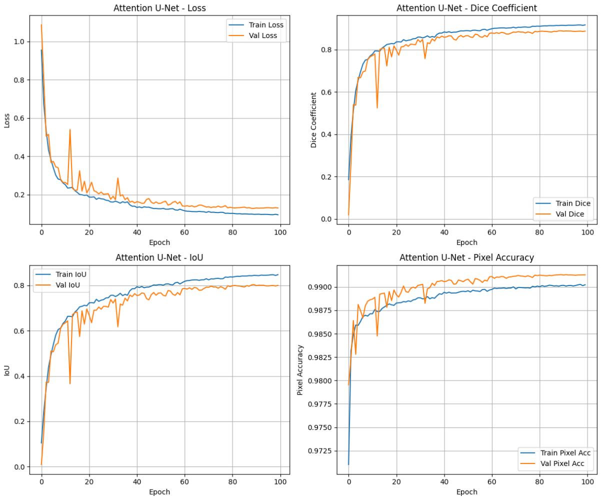

Fig. 3: Attention U-Net Training Curves

Figure 4 Explanation. This figure presents Joint U-Net training and validation curves for both segmentation (Dice/loss) and classification accuracy. Segmentation curves show mild late-epoch fluctuations, likely due to competing multi-task gradients, while classification accuracy remains high throughout, indicating effective joint optimization.

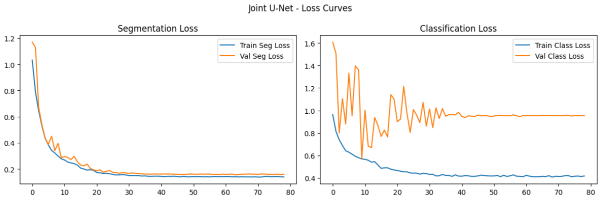

Fig. 4: Joint U-Net Training Curves

G. Aggregate Metric Comparisons

This section summarizes the endpoint performance of all models using bar charts for segmentation and classification metrics. These visualizations highlight the relative strengths and trade-offs of each approach.

Figure 5 Explanation. This bar chart compares mean Dice, IoU, and pixel accuracy for U-Net, Attention U-Net, and Joint U-Net on the test set. Error bars show standard deviation where available. Attention U-Net achieves the highest segmentation scores, while the joint model shows a slight trade-off, reflecting the effect of multi-task learning.

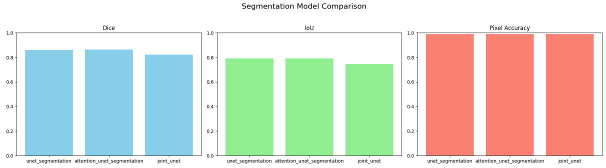

Fig. 5: Segmentation Bar Chart

Figure 6 Explanation. This bar chart displays accuracy, macro F1, and weighted F1 for the standalone classifier and joint model. Both models perform strongly, but the joint model has a minor reduction in macro F1, highlighting the impact of class imbalance and shared representation.

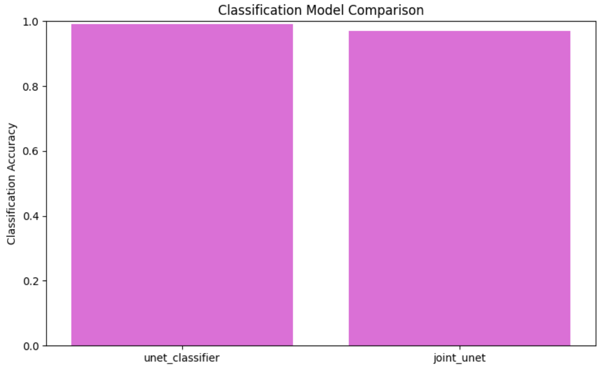

Fig. 6: Classification Bar Chart

H. Localized Views

This section provides qualitative visualizations of segmentation results, focusing on localized regions to illustrate the differences between U-Net and Attention U-Net predictions.

Figure 7 Explanation. This figure shows a localized view (LoU) for U-Net segmentation. The panels (left to right) display the original MRI, ground truth mask, and predicted mask overlay. Mild peripheral false positives are visible, especially at lesion boundaries, indicating some over-segmentation and less precise boundary adherence.

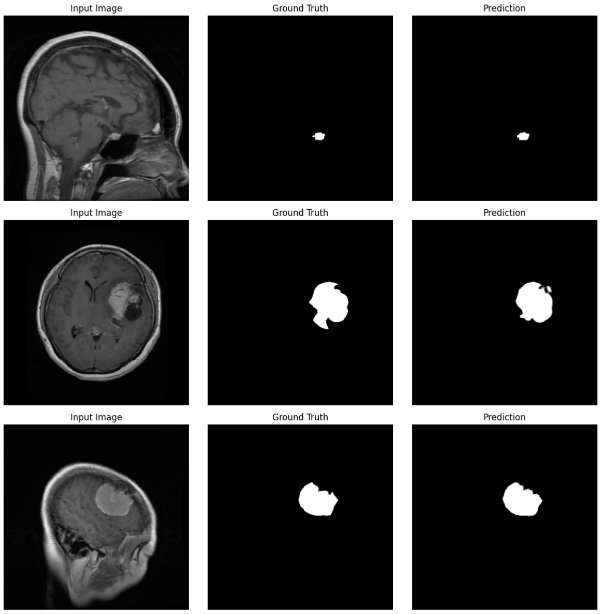

Fig. 7: Localized view (LoU) for U-Net segmentation.

Figure 8 Explanation. This figure presents a localized view (LoU) for Attention U-Net segmentation. The panels (left to right) show the original MRI, ground truth mask, and predicted mask overlay. Attention gating produces crisper lesion boundaries and reduces spurious activations, especially at the tumor periphery, demonstrating improved qualitative performance over the baseline.

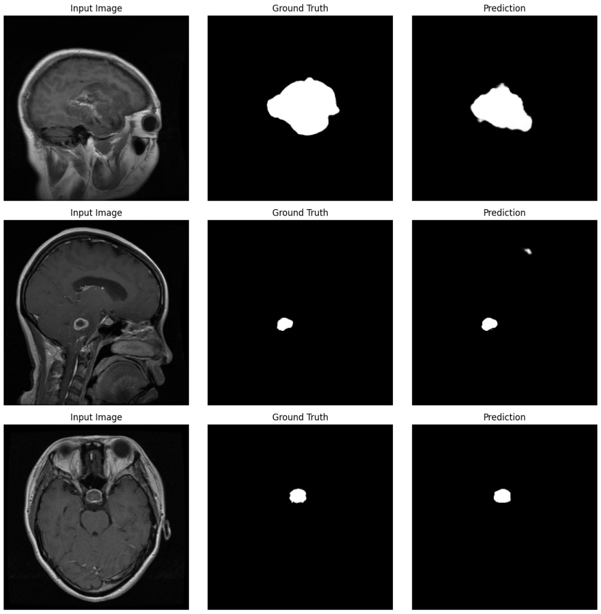

Fig. 8: Localized view (LoU) for AttU-Net segmentation.

V. Discussion

This study demonstrates that attention mechanisms, when integrated into encoder–decoder architectures, provide consistent but incremental improvements in boundary delineation for brain tumor segmentation, echoing findings in broader medical image analysis [5], [6]. The joint multi-task model achieves high classification accuracy with only a modest trade-off in segmentation performance, supporting the efficiency of shared representations for related tasks [3]. While gradient interference is observed, the overall discriminative power is preserved, and the unified approach offers practical deployment advantages (single checkpoint, unified preprocessing, reduced inference latency). These results align with recent advances in multi-task and attention-based deep learning [7].

VI. Conclusion

We present a reproducible multi-task deep learning pipeline for brain tumor MRI analysis, achieving robust segmentation and classification performance on the BRISC dataset [4]. The integration of attention modules and joint optimization not only improves boundary delineation and model efficiency but also demonstrates the feasibility of consolidating multiple diagnostic tasks into a single, deployable system. This approach reduces computational overhead and streamlines clinical workflows, making it attractive for real-world medical imaging applications. While minor trade-offs in segmentation accuracy are observed, the overall performance remains high, and the unified model offers a practical solution for comprehensive tumor assessment. Our results reinforce the value of unified models in medical imaging and provide a strong, extensible baseline for future research and clinical translation.

VII. Limitations

This work is subject to several limitations: (1) single-split evaluation without cross-validation, limiting statistical robustness; (2) reliance on 2D slice processing, which ignores volumetric continuity; (3) missing standard deviations for joint segmentation metrics; (4) lack of calibration and robustness analysis under noise or domain shift; and (5) absence of external institutional validation. Addressing these will be crucial for clinical translation.

VIII. Future Work

Future research should pursue multi-fold or nested validation, volumetric (3D/2.5D) or hybrid CNN–transformer backbones, adaptive loss balancing (e.g., uncertainty weighting, gradient surgery), test-time augmentation and ensembling, uncertainty and calibration assessment, and robustness stress testing under domain shifts. Incorporating recent advances in self-configuring segmentation frameworks [6] and large-scale transformer models [7] may further improve performance.

IX. Acknowledgement

The author gratefully acknowledges the support and coding assistance provided by peers and colleagues during this project. Their contributions were invaluable to the successful implementation and evaluation of the models.

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)*, pp. 234–241, Springer, 2015.

[2] O. Oktay, J. Schlemper, L. L. Folgoc, *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.

[3] R. Caruana, “Multitask learning,” *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[4] A. Fateh, Y. Rezvani, S. Moayedi, S. Rezvani, F. Fateh, and M. Fateh, “Brisc: Annotated dataset for brain tumor segmentation and classification with swin-hafnet,” *arXiv preprint arXiv:2506.14318*, 2025.

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[6] F. Isensee, P. F. Jaeger, S. 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] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.