Title

Evidential Deep Learning Ensembles for Uncertainty-Aware Segmentation and Volumetry of Diffuse Midline Gliomas

ISBN

9798190915129

Author

Kasi, Karthik

Publication Date

2026-09-09

Peer reviewed|Thesis/dissertation

Evidential Deep Learning Ensembles for Uncertainty-Aware Segmentation and Volumetry of Diffuse Midline Gliomas

by
Karthik Kasi

THESIS
Submitted in partial satisfaction of the requirements for degree of
MASTER OF SCIENCE

in

Biomedical Imaging

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Andreas Rauschecker
704B2305CFB3407...

Andreas Rauschecker
Chair

DocuSigned by:
Alexandra Gersing
[illegible]45C...

Alexandra Gersing

DocuSigned by:
Renuka Sriram
[illegible]07...

Renuka Sriram

DocuSigned by:
Yang Yang
731E1D96DE5B471...

Yang Yang

Committee Members

Dedication and Acknowledgements

I would like to thank Atlas Haddadi, Pierre Nedelec, and Emma Duprat for their guidance, collaboration, and support throughout this project. I am also grateful to Yassine Guennoun for his initial work on the model and for the documentation that helped provide a foundation for this work. Finally, I would like to express my sincere gratitude to Andreas Rauschecker for giving me the opportunity to join his lab and for being an exceptional mentor throughout this project. His guidance, encouragement, and support have been invaluable to both this work and my development as a researcher.

Abstract

Evidential Deep Learning Ensembles for Uncertainty-Aware Segmentation and Volumetry of Diffuse Midline Gliomas

By: Karthik Kasi

Diffuse midline gliomas (DMGs) have infiltrative, poorly defined boundaries that make tumor measurement difficult and variable. Volumetric assessment may be more informative than bidimensional response measurements, but manual volumetry is time-consuming and automated methods typically provide a single volume without indicating reliability. We evaluated whether evidential deep learning (EDL) with model ensembling could generate calibrated 95% uncertainty intervals for automated DMG volumes.

Two five-model EDL ensembles based on SegResNet and nnU-Net v2 were evaluated. Voxel-wise probability distributions were propagated into tumor-volume estimates and 95% intervals. Analyses included voxel-wise calibration, true-volume coverage, relative interval width, and comparison of SegResNet uncertainty with physician difficulty ratings and spatial uncertainty annotations.

The EDL nnU-Net v2 Ensemble approached 95% coverage but produced intervals likely too broad for clinical use, whereas the EDL SegResNet Ensemble provided the strongest balance between coverage and precision. SegResNet intervals widened with physician-rated difficulty, while aleatoric and epistemic uncertainty localized near physician-identified uncertain regions and tumor boundaries. These findings suggest evidential ensembling can provide automated DMG volumetry with calibrated uncertainty estimates that reflect clinically recognizable sources of difficulty.

Table of Contents

List of Figures

Introduction

Diffuse midline gliomas are aggressive pediatric tumors that infiltrate critical midline brain structures and carry a poor prognosis (Pachocki and Hol, 2022). Progression and treatment response are assessed primarily on serial MRI (Gilligan et al., 2020; Warren et al., 2013). Current Response Assessment in Pediatric Neuro-Oncology criteria rely largely on bidimensional measurements, yet volumetric tumor burden may be more sensitive to clinically meaningful change and more strongly associated with outcome (Gilligan et al., 2020; Khalili et al., 2024). Accurate, reproducible segmentation is therefore central to longitudinal volumetry.

DMGs are particularly difficult to segment because their diffuse, infiltrative growth obscures the boundary between tumor and normal tissue, resulting in substantial inter-segmenter variability (Pachocki and Hol, 2022). Manual contouring is therefore labor-intensive and inconsistent, further complicated by heterogeneous MRI signal and indistinct tumor–tissue interfaces. Deep learning offers rapid, reproducible segmentation, with performance approaching expert readers in recent multi-dataset studies (Boyd et al., 2024; Kazerooni, 2024). However, most models reduce voxel-wise probabilities to a single binarized segmentation (Mehrtash et al., 2020). If those probabilities are uncalibrated or overconfident, the contour can conceal substantial uncertainty, particularly in atypical or difficult scans (Mehrtash et al., 2020).

A calibrated interval around tumor volume could make automated measurements clinically actionable. A narrow interval would support confidence that the reported volume is reliable, whereas a wide interval would signal that the scan requires expert review. Across serial MRI, these intervals could show whether an apparent increase or decrease exceeds segmentation uncertainty, helping distinguish likely biological change from measurement variability. Without

calibration, clinicians cannot know when to trust an automated volume or when uncertainty could alter interpretation.

Meaningful intervals must account for two sources of uncertainty. Epistemic uncertainty reflects differences in learned knowledge or methodology (Huang et al., 2022; Zepf, 2024). Radiologists may contour differently because they follow different conventions—such as whether to include necrotic tissue—or interpret findings through different experience; models can likewise learn different rules from available annotations. Clearer labeling standards, more representative annotations, or improved training may reduce this uncertainty (Huang et al., 2022; Zepf, 2024). Aleatoric uncertainty reflects ambiguity inherent in the image (Huang et al., 2022; Wang et al., 2019). Even an experienced radiologist using a consistent method may remain unsure of a boundary because of low contrast, artifacts, partial-volume effects, or infiltrative growth (Wang et al., 2019).

Deep ensembles capture epistemic uncertainty by treating independently trained networks as a panel of computational readers. Agreement supports a reliable segmentation; disagreement identifies predictions that depend strongly on learned parameters. This variation can flag unfamiliar scans and define a distribution of plausible tumor volumes rather than a single estimate.

Evidential deep learning complements this approach by representing uncertainty within each network. Standard models produce a point estimate of class probability—typically through a sigmoid or softmax—and threshold it into a binary mask, indicating the preferred label but not its evidential support. EDL instead learns non-negative evidence for each class and uses that evidence to parameterize a distribution over possible class probabilities (Hu et al., 2024; Li et al., 2022). Common implementations use a Beta distribution for binary segmentation or a Dirichlet

distribution for multiclass segmentation (Tan et al., 2024). The distribution's mean gives the predicted probability, while its concentration reflects evidential strength: clear features yield concentrated evidence, whereas ambiguous boundaries yield weaker, broader evidence (Hu et al., 2024; Tan et al., 2024). EDL thus preserves voxel-level uncertainty that a single probability and binarized contour can obscure (Li et al., 2022).

The approaches are complementary: ensembles capture variation across learned models, while EDL captures uncertainty in each voxel-level prediction. Together, they can produce volume intervals that reflect uncertainty in both the model and image.

This thesis evaluates uncertainty-aware approaches to DMG segmentation and volumetry. Two five-model EDL ensembles—the lab-developed EDL SegResNet Ensemble (Guennoun et al., 2026) and a newly implemented EDL nnU-Net v2 Ensemble—were compared with a standard nnU-Net v2 benchmark. Performance was assessed using segmentation accuracy, voxel-wise probability calibration, tumor-volume interval coverage, and relative interval width. The EDL SegResNet Ensemble was further evaluated against assessments from an independent physician to determine whether model performance and interval width reflected image difficulty and whether model-derived aleatoric and epistemic uncertainty localized to physician-identified regions of uncertainty. Together, these analyses examined whether the models could preserve segmentation accuracy while providing calibrated, informative estimates of uncertainty in tumor segmentation and volume.

Methods

Model Architectures and Inference

EDL SegResNet Ensemble

The first model was the EDL SegResNet Ensemble, a homogeneous ensemble of five SegResNet-based EDL models adapted from the prior meningioma uncertainty framework (Guennoun et al., 2026). Each model was initialized from the meningioma-pretrained network and fine-tuned on DMG FLAIR images with all layers unfrozen, allowing both the learned features and segmentation head to adapt to the new task. This transfer-learning strategy was used because it outperformed training the same architecture from scratch.

EDL nnU-Net v2 Ensemble

The second model was the EDL nnU-Net v2 Ensemble, which used the same five-member evidential ensemble structure as the EDL SegResNet Ensemble but replaced the EDL SegResNet backbones with nnU-Net v2 models modified to produce evidential outputs (Guennoun et al., 2026). In each ensemble member, nnU-Net v2 served as the base segmentation architecture, with the output adapted to estimate non-negative evidence for tumor and non-tumor classes rather than only a single deterministic probability map. All five models were trained independently from scratch on the DMG training cohort. This model tested whether the evidential ensemble framework generalized to a self-configuring segmentation architecture.

Standard nnU-Net v2 Control

The third model was a standard nnU-Net v2 binary segmentation network trained from scratch. Because nnU-Net v2 is a widely used modern segmentation framework, it served as a benchmark for comparing the performance and interval-based outputs of the two EDL ensembles (Isensee et al., 2020).

Inference

At inference, the EDL SegResNet Ensemble and EDL nnU-Net v2 Ensemble each produced a voxel-wise Beta distribution for tumor probability, following the framework described by (Guennoun et al., 2026). The 2.5th-percentile, 50th-percentile, and 97.5th-percentile probability maps were extracted and thresholded at 0.5 to generate inclusive, median, and conservative binary segmentations, respectively. Tumor volumes were calculated from each segmentation. The median segmentation provided the point estimate of tumor volume, while the volumes derived from the 2.5th- and 97.5th-percentile segmentations defined the bounds of the 95% interval.

Standard nnU-Net v2 control. The standard nnU-Net v2 produced a single voxel-wise tumor-probability map. Probabilities were thresholded at 0.5 to obtain the binary segmentation used for the point estimate of tumor volume (median segmentation). For comparison with the EDL-derived interval, voxels with tumor probabilities greater than or equal to 0.025 were included in an inclusive segmentation, whereas voxels with probabilities greater than or equal to 0.975 were retained in a conservative segmentation. The corresponding volumes were used as bounds around the standard model's point estimate.

Training and Testing Datasets

Training dataset

The training cohort comprised 591 FLAIR MRI images collected from four institutions or datasets: the University of Zurich, the University of California, San Francisco, the University of Michigan, and the public Brain Tumor Segmentation in Pediatrics (BraTS-PEDs) dataset (Fathi Kazerooni et al., 2025). Patients contributed an average of 2.13 images each. Of the 591 images, 177 were confirmed pretreatment, 329 were confirmed post-treatment, and 85 had unknown treatment status. BraTS-PEDs images were not labeled as pre- or post-treatment. Each image was accompanied by a physician-defined tumor segmentation that served as the ground truth. Images that were not already skull stripped underwent brain extraction using HD-BET (Isensee et al., 2019). For both EDL ensembles, each of the five constituent models received a different 80% subset of the training cohort to promote diversity across ensemble members. The EDL SegResNet Ensemble members were fine-tuned on these subsets, whereas the EDL nnU-Net v2 Ensemble members were trained from scratch. The standard nnU-Net v2 control was trained from scratch using 100% of the training cohort.

Testing dataset

The independent test cohort comprised 89 FLAIR MRI images: 15 held-out images from BraTS-PEDs and 74 images from the Pacific Pediatric Neuro-Oncology Consortium (PNOC). Patients contributed an average of 1.51 images each. Of the 89 images, 25 were confirmed pretreatment, 49 were confirmed post-treatment, and 15 had unknown treatment status. All testing images had a physician-defined tumor segmentation, which served as the ground truth for

evaluation. Images that were not already skull stripped underwent brain extraction using HD-BET (Isensee et al., 2019).

Cross-model Comparison Metrics

Dice Similarity Coefficient

Segmentation performance was assessed using the standard Dice similarity coefficient, which measures the spatial overlap between the predicted and ground-truth segmentations. For both EDL ensembles, Dice was calculated only from the median binary segmentation. For the standard nnU-Net v2 control, Dice was calculated from the binary segmentation obtained by thresholding the tumor-probability map at 0.5.

Volume Coverage

Volume coverage measured the percentage of test examinations for which the ground-truth tumor volume fell within the model's predicted lower and upper volume bounds. Evaluation focused on whether each interval-generation method achieved the intended 95% coverage target. Lower coverage indicated that intervals excluded the ground truth too often, whereas substantially higher coverage could reflect overly conservative intervals with limited clinical precision.

Voxel-wise Probability Calibration

Voxel-wise calibration assessed whether predicted tumor probabilities matched observed tumor frequencies across the test set. Voxels were grouped into 10 equal-width probability bins spanning 0 to 1. Within each bin, the mean predicted tumor probability was compared with the proportion of voxels labeled as tumor in the ground-truth segmentation. Thus, a bin with a mean

predicted probability of approximately 0.8 was well calibrated if approximately 80% of its voxels were tumor. These quantities were first accumulated for each test image. For the pooled calibration curve, voxel counts, sums of predicted probabilities, and ground-truth tumor counts were then summed across the test set before the bin-level means and observed frequencies were calculated. This analysis evaluated calibration of voxel-level probabilities and was distinct from volume coverage, which tested whether the ground-truth tumor volume fell within predicted volume intervals at specified nominal coverage levels.

Relative Interval Width

High coverage can be achieved simply by generating very wide volume intervals, but such intervals may be too imprecise to guide clinical interpretation. Relative interval width was therefore used to assess interval tightness. For each examination, the difference between the upper and lower predicted volumes was normalized by the ground-truth tumor volume and expressed as a percentage, allowing interval size to be interpreted relative to tumor burden. Smaller values indicate more precise intervals, whereas larger values indicate greater uncertainty and reduced clinical usefulness. Relative interval width was considered alongside coverage to identify models that approached the 95% coverage target without producing unnecessarily broad intervals.

Physician–Model Uncertainty Concordance

Physician Difficulty and Uncertainty Annotations

A physician who was not involved in generating the ground-truth segmentations or training the models independently rated the segmentation difficulty of every test image on a five-point scale, where 1 indicated low difficulty and 5 high difficulty. For a selected subset of test

images and slices, the physician also labeled voxels as confidently tumor, confidently non-tumor, or confusing. Voxels labeled as confusing formed the physician reference map of spatial uncertainty.

Model Uncertainty Analysis and Physician Concordance

Corresponding aleatoric and epistemic uncertainty maps were generated from the EDL SegResNet Ensemble. Epistemic uncertainty reflected concordance across the five ensemble members, whereas aleatoric uncertainty reflected the width of each model's voxel-wise output distribution; exact derivations followed (Guennoun et al., 2026). For each annotated image, the number of voxels in the physician's confusing-region map was counted. Each model uncertainty map was then thresholded to retain the same number of highest-uncertainty voxels, producing matched binary maps for comparison. Physician-defined confusing regions served as the reference. Spatial agreement with the aleatoric and epistemic maps was assessed visually and quantified using the Dice similarity coefficient and average symmetric surface distance (ASSD).

Uncertainty was also characterized relative to the ground-truth tumor boundary for each test image. Signed voxel distances were negative inside the tumor, positive outside the tumor, and 0 mm at the boundary. Voxels were grouped into distance bins, and mean aleatoric and epistemic uncertainty was calculated within each bin for each image. Image-level values were summarized across the test set by the median and 2.5th–97.5th percentile range. To compare spatial patterns rather than absolute scale, each uncertainty measure was normalized to its own maximum median value.

Results

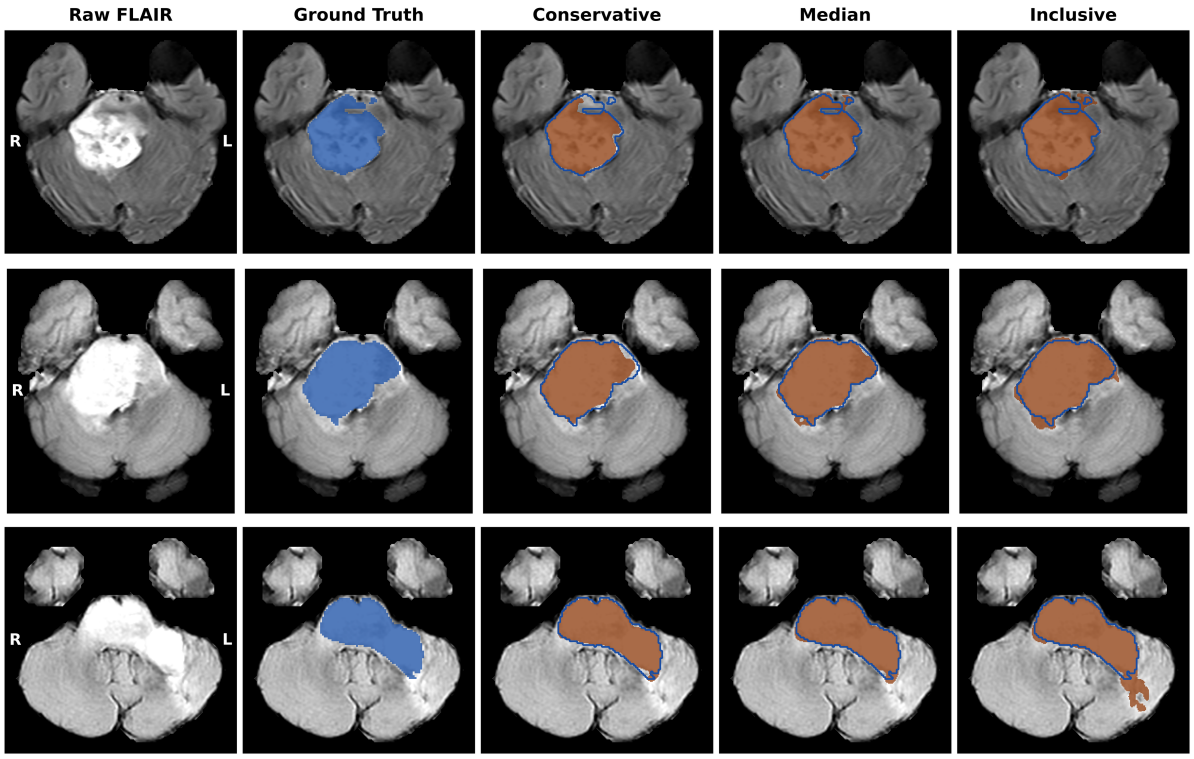

Figure 1. Example EDL SegResNet segmentations.

Three representative test-set cases are shown. Columns display the raw skull-stripped FLAIR image, corrected ground-truth tumor segmentation, and conservative, median, and inclusive EDL SegResNet segmentations. The conservative, median, and inclusive masks represent progressively less restrictive uncertainty thresholds. Predicted segmentations (orange) are shown as colored overlays, with the ground-truth boundary outlined (blue) for comparison.

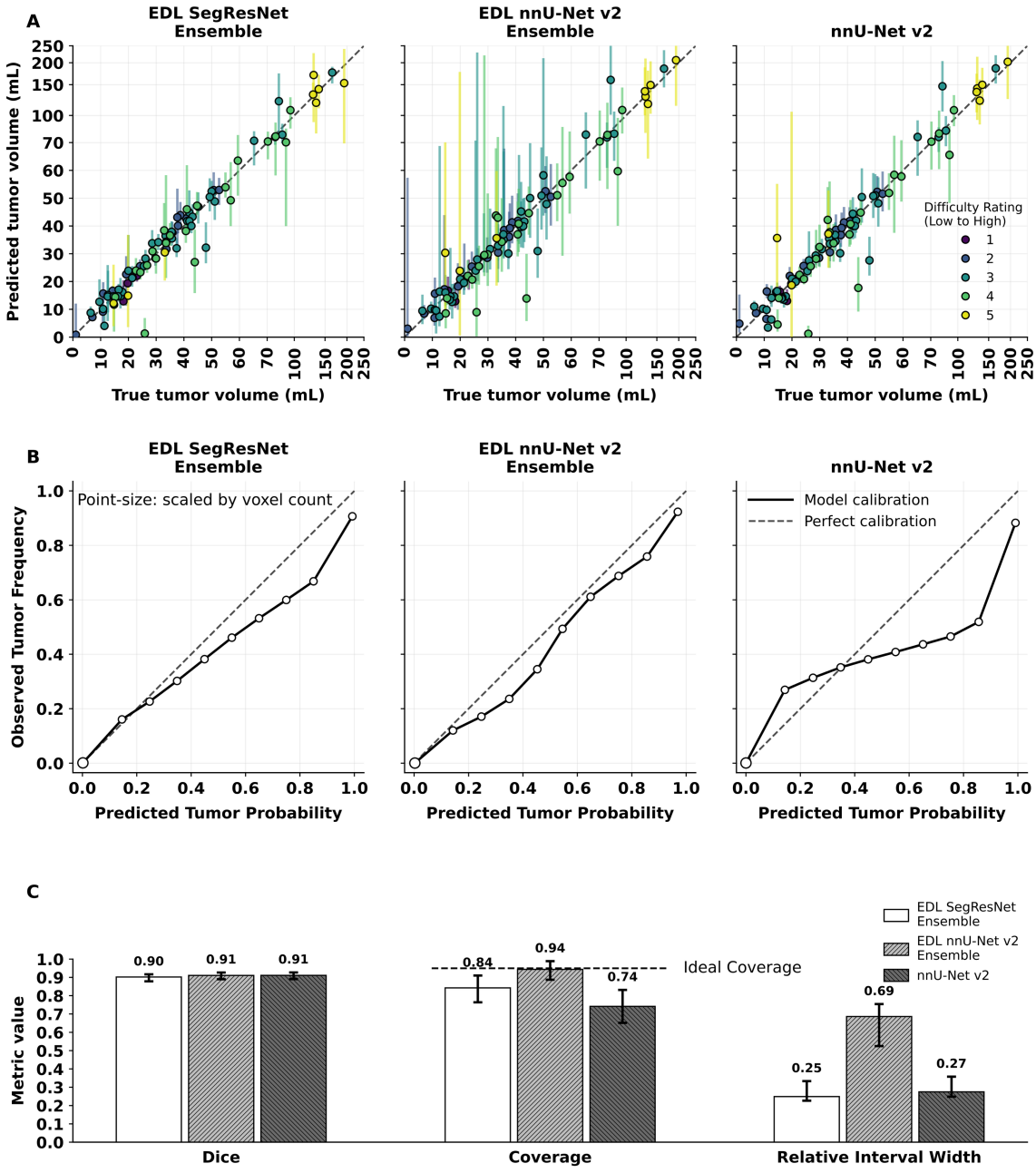

Figure 2. Cross-Model Performance.

(A) True versus predicted tumor volume on the test set for the EDL SegResNet Ensemble, EDL nnU-Net v2 Ensemble, and standard nnU-Net v2. Points represent individual test images and are colored by physician-assigned difficulty rating (1–5, low to high). Vertical error bars show each image's 95% tumor-volume interval, and the dashed diagonal line indicates perfect agreement. **(B)** Voxel-wise calibration on the test set, showing mean predicted tumor probability versus observed tumor frequency across pooled probability bins. Marker size reflects the number of voxels in each bin, and the dashed diagonal line indicates perfect calibration. **(C)** Test-set Dice, 95% volume coverage, and relative interval width across models. Values represent median image-level Dice, the proportion of test images covered by the 95% tumor-volume interval, and median image-level relative interval width. Error bars show bootstrap 95% confidence intervals from 20,000 resamples, and the dashed horizontal line marks nominal 95% coverage.

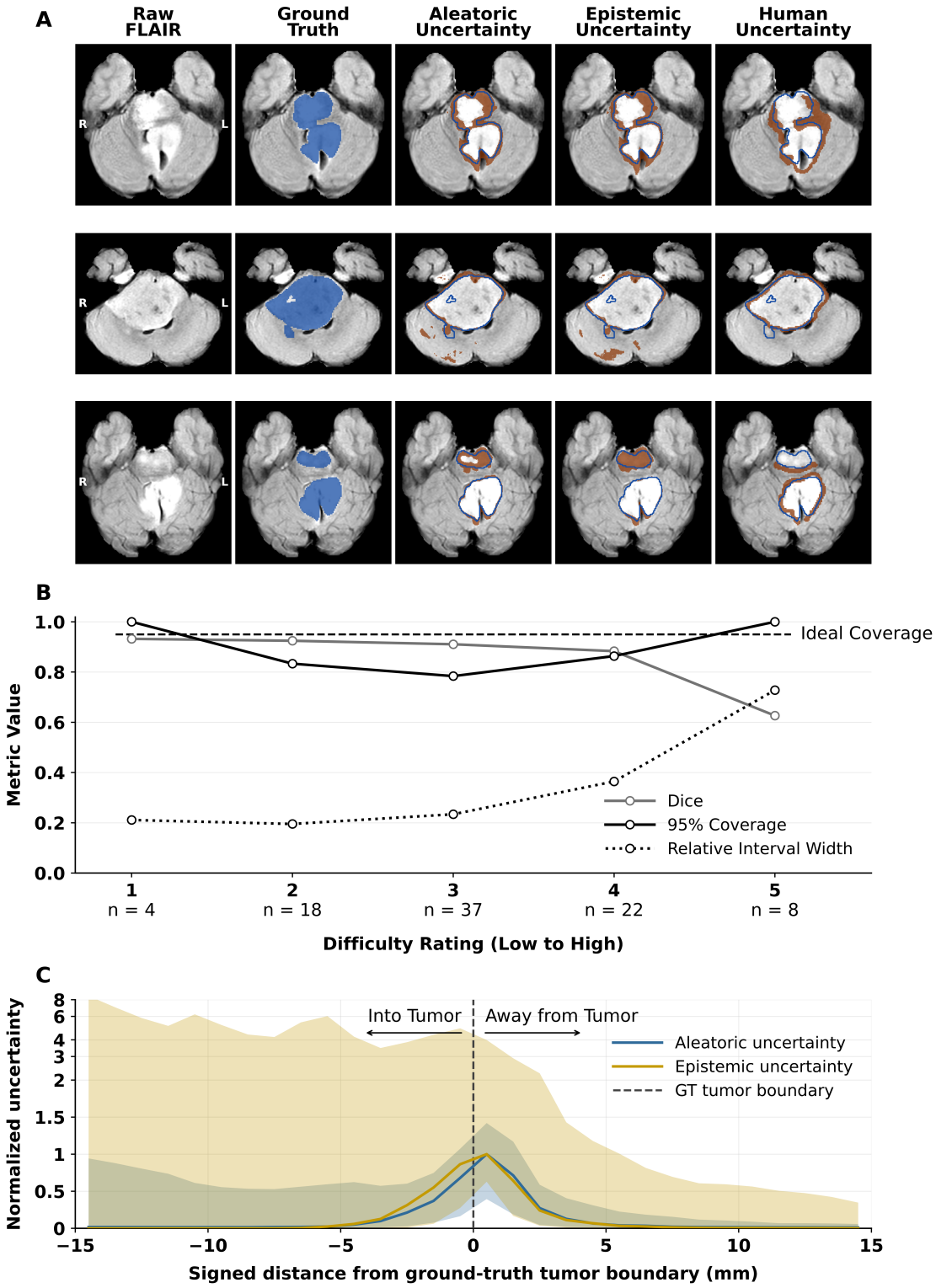

Figure 3. Physician Alignment and Spatial Characteristics of EDL SegResNet Ensemble Uncertainty.

(A) Three representative test-set cases showing raw FLAIR, ground-truth tumor segmentation, aleatoric uncertainty, epistemic uncertainty, and physician-annotated uncertainty. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) The uncertainty annotations were provided by an independent physician who was not involved in creating the ground-truth segmentations. For each slice, the number of voxels marked as uncertain by the physician was counted, and the same number of voxels with the highest aleatoric or epistemic uncertainty values were selected for comparison. Ground-truth tumor boundaries are outlined. **(B)** EDL SegResNet Ensemble Dice, 95% volume coverage, and relative interval width across physician-assigned test-set difficulty ratings from 1–5. Values represent the median Dice, proportion of cases covered by the 95% volume interval, and median relative interval width within each difficulty group. The dashed line marks nominal 95% coverage, and n indicates the number of test-set cases in each group. **(C)** Aleatoric and epistemic uncertainty as a function of signed distance from the ground-truth tumor boundary across test-set cases. Voxels were grouped into signed-distance bins and uncertainty was averaged within each bin for each case. Solid lines show the median across cases, with shaded regions spanning the 2.5th–97.5th percentiles. Negative distances extend into the tumor, and positive distances extend away from the tumor; the dashed vertical line at x = 0 marks the ground-truth tumor boundary.

As expected, estimated tumor burden increased progressively from the conservative to the median and inclusive segmentations. Visual inspection showed that these increases occurred primarily along the tumor margins, where uncertainty in the extent of diffuse disease was greatest, rather than through large changes within the tumor core (Figure 1).

Median 95% volume coverage differed across models: 84% for the EDL SegResNet Ensemble, 94% for the EDL nnU-Net v2 Ensemble, and 74% for standard nnU-Net v2 (Figure 2A, C). Although the EDL nnU-Net v2 Ensemble came closest to the nominal 95% target, this coverage was achieved with a median relative interval width of 69%, producing intervals that were likely too broad for practical clinical use. The EDL SegResNet Ensemble and standard nnU-Net v2 produced substantially tighter intervals, with median relative interval widths of 25% and 27%, respectively. Despite these differences in interval performance, segmentation accuracy was similar across models: median Dice was 0.90 for the EDL SegResNet Ensemble and 0.91 for both the EDL nnU-Net v2 Ensemble and standard nnU-Net v2 (Figure 2C). For all 3 models, Dice was calculated from the median segmentation, as described in Methods.

Both EDL ensembles showed markedly better voxel-wise calibration than standard nnU-Net v2, with predicted tumor probabilities more closely matching observed tumor frequencies (Figure 2B). The EDL models also produced a broader range of predicted probabilities,

indicating greater differentiation among low-, intermediate-, and high-confidence voxel predictions.

Physician-annotated uncertainty visually overlapped with both aleatoric and epistemic uncertainty from the EDL SegResNet Ensemble, with all three concentrated near the tumor boundary (Figure 3A, C). As expected, Dice was lower for images assigned the highest difficulty ratings of 4 and 5 (Figure 3B). Relative interval width also increased in these groups, indicating greater model uncertainty for more difficult images. Volume coverage was highest for difficulty ratings 1 and 5 and showed a modest decrease across the intermediate ratings of 2–4, but remained relatively consistent overall (Figure 3B).

Quantitatively, physician and model uncertainty showed limited exact voxel-wise overlap: Dice was 0.43 for aleatoric uncertainty and 0.41 for epistemic uncertainty. However, their spatial localization was close, with low average symmetric surface distances of 1.96 mm and 2.21 mm, respectively. Thus, although the model did not reproduce the physician's uncertainty annotations voxel for voxel, it consistently identified nearby regions of uncertainty (Figure 3A).

Discussion

Among the three models, the EDL SegResNet Ensemble provided the strongest balance between volume coverage and interval width. The EDL nnU-Net v2 Ensemble approached the nominal 95% coverage target, but its substantially wider intervals limited their practical value. Its weaker balance may reflect training from scratch without the benefit of pretraining, or a mismatch between the evidential objective and nnU-Net v2's self-configuring pipeline, which is optimized primarily for segmentation accuracy rather than uncertainty estimation. These

explanations require direct testing. Importantly, all three models achieved similar Dice scores, suggesting that uncertainty-aware ensembling did not materially reduce point-estimate segmentation performance despite the additional computational demands of training multiple models (Mehrtash et al., 2020).

The markedly better voxel-wise calibration of both EDL ensembles is also important because it makes the predicted tumor probabilities more meaningful at each spatial location. This may be particularly valuable for boundary-sensitive applications such as radiation therapy or high-intensity focused ultrasound planning, where uncertainty about whether a voxel contains tumor could affect target delineation and the balance between treating infiltrative disease and sparing adjacent healthy tissue. Calibrated voxel-wise probabilities could identify regions that warrant closer review or more conservative planning.

Aleatoric and epistemic uncertainty showed broadly similar spatial patterns, with both concentrated near the tumor boundary. This overlap may reflect the diffuse biology of DMG: intrinsically indistinct margins can generate ambiguity within each model while also producing disagreement across models. The modest training-set size may have further limited the diversity of learned solutions, reducing separation between the two uncertainty signals. In this setting, the distinction between aleatoric and epistemic uncertainty may therefore be more useful conceptually than spatially.

The EDL SegResNet Ensemble performed worst on the same images the physician rated as most difficult and appropriately widened its volume intervals for those images. This behavior strengthens the model's trustworthiness: rather than remaining confidently wrong when the task became difficult, it communicated reduced certainty and exposed measurements that warrant greater caution or expert review. Model-derived uncertainty was also localized near physician-

identified confusing regions. Although exact voxel-wise overlap was modest, the low surface distances indicate that the model and physician identified uncertainty in closely neighboring regions. These findings do not establish that the model reasons like a physician, but they suggest that its uncertainty is responsive to clinically recognizable sources of difficulty rather than to an entirely model-specific failure pattern. Such alignment may make model outputs easier to interpret and increase clinical trustworthiness.

Limitations and Future Directions

This study was limited by the size of the training and test cohorts, which reduced bootstrap precision and precluded firm conclusions about many between-model differences. This constraint is common in DMG research because the disease is rare and aggressive, but it remains important for interpreting the estimates. More fundamentally, each image was paired with a single binary physician segmentation. Training an evidential model to represent uncertainty from a deterministic label is an inherently limited objective: legitimate boundary disagreement is collapsed into a single reference, and the model is penalized for expressing alternatives that may be clinically plausible. Future datasets should include independent segmentations from multiple physicians and preserve their degree of agreement as supervision—for example, encoding a voxel as strongly supported when all readers label it tumor and as uncertain when readers disagree. This would allow evidential outputs to learn observed inter-segmenter variability directly rather than infer it from binary labels.

Conclusion

This study demonstrates that evidential deep-learning ensembles can preserve high DMG segmentation accuracy while providing information that a single binarized segmentation cannot: how reliable the predicted tumor probabilities and volumes are. Both EDL ensembles improved voxel-wise calibration relative to standard nnU-Net v2, and the EDL SegResNet Ensemble achieved the most favorable balance between volume coverage and interval width. Crucially, its intervals widened on images rated most difficult by the physician, and its uncertainty localized near physician-identified confusing regions and the diffuse tumor boundary. These findings strengthen the trustworthiness of the model because it signals when and where its predictions are less reliable rather than presenting every automated volume with equal confidence. Although the EDL nnU-Net v2 Ensemble showed that high coverage alone can produce intervals too broad for practical use, the EDL SegResNet Ensemble illustrates the potential of calibrated evidential modeling to support targeted review, more transparent longitudinal volumetry, and boundary-sensitive treatment planning. Overall, these findings establish EDL ensembles as a promising approach for making automated DMG volumetry more informative and trustworthy.

References

Boyd A, Ye Z, Prabhu SP, et al. Stepwise Transfer Learning for Expert-Level Pediatric Brain Tumor MRI Segmentation in a Limited Data Scenario. Research Publications (Maastricht University) 2024;6(4); doi: 10.1148/ryai.230254.

Fathi Kazerooni A, Khalili N, Liu X, Jiang Z, Gandhi D, Pavaine J, Shah LM, Jones BV, Sheth N, Prabhu SP, McAllister AS, Tu W, Nandolia KK, Rodriguez A, Shaikh IS, Sanchez-Montano M, Lai HA, Haldar D, Anderson H, Bagheri S, Zapaishchykova A, Familiar AM, Gottipati A, Haldar S, Khalili N, Maleki N, Poussaint T, Storm PB, Bornhorst M, Packer R, Hummel T, de Blank P, Hoffman L, Aboian M, Nabavizadeh A, Ware JB, Kann BH, Bakas S, Rood B, Resnick A, Vossough A and Linguraru MG. The Brain Tumor Segmentation in Pediatric Magnetic Resonance Imaging (BraTS-PEDs) (Version 1) [Data set]. The Cancer Imaging Archive 2025; doi: 10.7937/DX5C-TJ86.

Gilligan LA, DeWire-Schottmiller M, Fouladi M, et al. Tumor Response Assessment in Diffuse Intrinsic Pontine Glioma: Comparison of Semiautomated Volumetric, Semiautomated Linear, and Manual Linear Tumor Measurement Strategies. American Journal of Neuroradiology 2020;41(5):866–873; doi: 10.3174/ajnr.a6555.

Guennoun Y, Nédélec P, McArthur M, et al. Segmenting with Confidence: Uncertainty Quantification for Brain Tumor Imaging. Research Square 2026; doi: 10.21203/rs.3.rs-8407421/v1.

Hu C, Xia T, Cui Y, et al. Trustworthy Multi-Phase Liver Tumor Segmentation via Evidence-Based Uncertainty. Engineering Applications of Artificial Intelligence 2024;133:108289; doi: 10.1016/j.engappai.2024.108289.

Huang L, Ruan S and Denœux T. Application of Belief Functions to Medical Image Segmentation: A Review. HAL (Le Centre pour la Communication Scientifique Directe) 2022;91:737–756; doi: 10.1016/j.inffus.2022.11.008.

Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. Nature Methods 2020;18(2):203–211; doi: 10.1038/s41592-020-01008-z.

Isensee F, Schell M, Pflueger I, et al. Automated Brain Extraction of Multisequence MRI Using Artificial Neural Networks. Human Brain Mapping 2019;40(17):4952–4964; doi: 10.1002/hbm.24750.

Kazerooni AF. IMG-12. CBTN BRAIN TUMOR SEGMENTATION INITIATIVE: UPDATES ON MODEL RELEASE, HGG SEGMENTATION, AND SURVIVAL ANALYSIS. Neuro-Oncology 2024;26:0; doi: 10.1093/neuonc/noae064.349.

Khalili N, Khalili N, Gandhi D, et al. NIMG-70. COMPARISON OF VOLUMETRIC AND CROSS-SECTIONAL MEASUREMENTS IN DIFFUSE MIDLINE GLIOMAS: RETHINKING RESPONSE ASSESSMENT. PubMed Central 2024;26; doi: 10.1093/neuonc/noae165.0834.

Li H, Yang N, Ser JD, et al. Region-Based Evidential Deep Learning to Quantify Uncertainty and Improve Robustness of Brain Tumor Segmentation. Neural Computing and Applications 2022;35(30):22071–22085; doi: 10.1007/s00521-022-08016-4.

Mehrtash A, Wells WM, Tempany CM, et al. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. arXiv (Cornell University) 2020;39(12):3868–3878; doi: 10.1109/tmi.2020.3006437.

Pachocki CJ and Hol EM. Current Perspectives on Diffuse Midline Glioma and a Different Role for the Immune Microenvironment Compared to Glioblastoma. Journal of Neuroinflammation 2022;19(1):276; doi: 10.1186/s12974-022-02630-8.

Tan HS, Wang K and McBeth R. Uncertainty-Error Correlations in Evidential Deep Learning Models for Biomedical Segmentation. arXiv (Cornell University) 2024; doi: 10.48550/arxiv.2410.18461.

Wang G, Li W, Aertsen M, et al. Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks. Neurocomputing 2019;338:34–45; doi: 10.1016/j.neucom.2019.01.103.

Warren KE, Poussaint TY, Vézina G, et al. Challenges with Defining Response to Antitumor Agents in Pediatric Neuro-Oncology: A Report from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) Working Group. Pediatric Blood & Cancer 2013;60(9):1397–1401; doi: 10.1002/pbc.24562.

Zepf K. Aleatoric and Epistemic Uncertainty in Image Segmentation. Technical University of Denmark, DTU Orbit (Technical University of Denmark, DTU) 2024.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

Signed by:
Kartkik Kasi
57CC0B258134492...
Author Signature

08/30/2026
Date