SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Beyond Binary Masks: Stochastic Ensembles for Uncertainty-Aware Tumor Segmentation

Andres Guerrero, UCI Center for Applied Artificial Intelligence Research; Peter Chang, MD

Introduction/Background

Accurate tumor segmentation is essential for cancer diagnosis. This task often relies on manually annotated datasets, where regions are labeled using binary masks. However, these hard labels do not capture the uncertainty in tumor boundaries. This study shows that by mimicking the variability seen between and within human annotators, we can introduce controlled ambiguity into segmentation models. This approach approximates the inherent uncertainty in tumor segmentation, thereby improving the interpretability and performance of model predictions compared to standard deterministic methods.

Methods/Intervention

To simulate interobserver variability, we train three U-Net–like convolutional neural networks on the BraTS-GoAT 2024 dataset, each using a different patch size to vary the level of contextual information available. This yields three models with differing segmentation capabilities. Intraobserver variability is emulated by injecting Gaussian noise into the input images prior to inference. Each model performs inference on ten noise-perturbed versions of the same input, resulting in a total of thirty predictions per case. The final ensemble output is obtained by aggregating all thirty logits maps. Compared to the deterministic ensemble produced by the same three models without input perturbation, this stochastic approach generates a distribution over the logits space that more accurately reflects regional uncertainty (Figure 1). This probabilistic representation enables per-blob confidence estimation in the segmentation masks (Figure 2), allowing for more robust and interpretable thresholding.

Results/Outcome

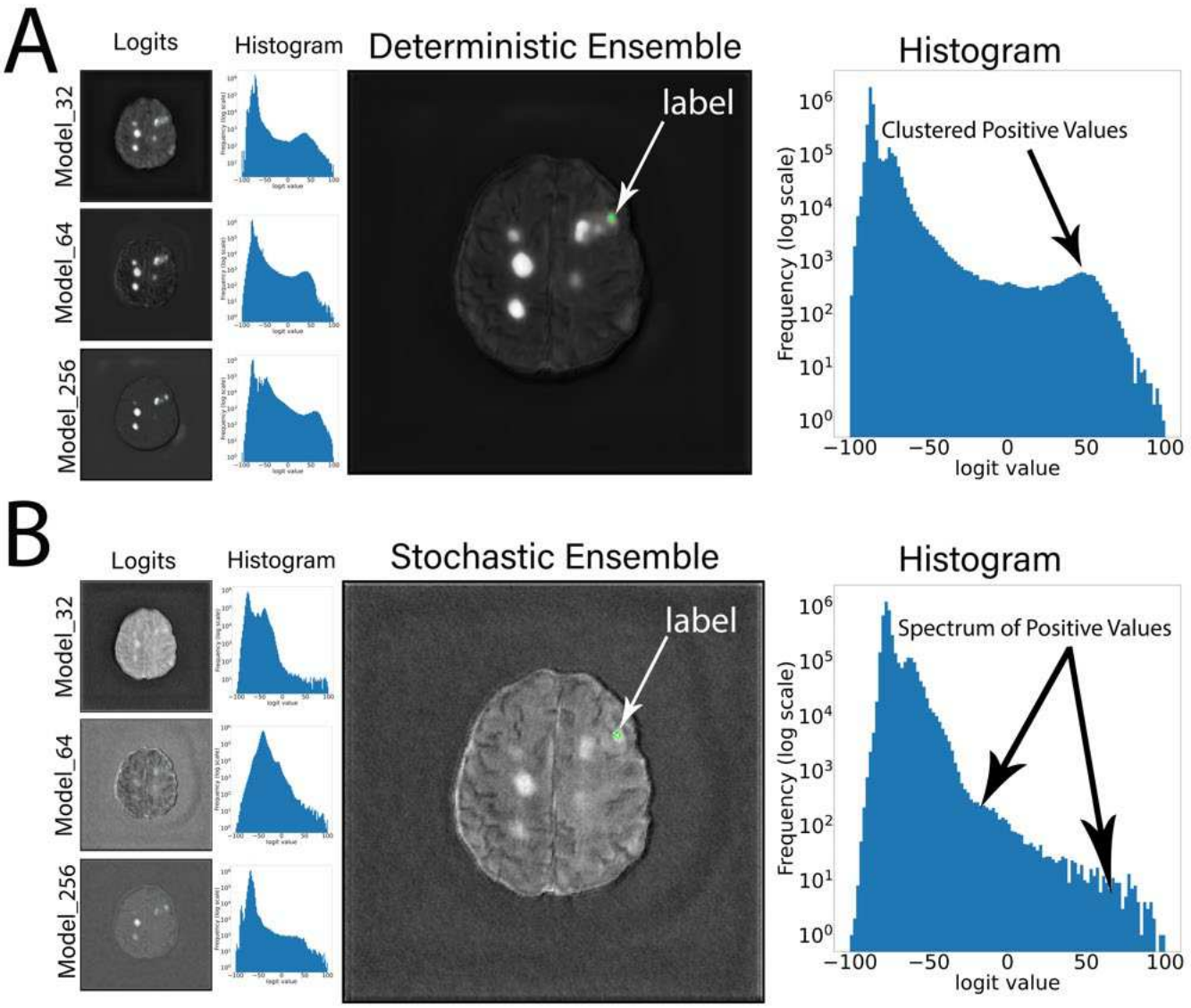
In a validation cohort of 100 patients, the logit distributions produced by the stochastic and deterministic ensembles exhibited notable differences. The stochastic ensemble yielded a lower mean logit value (27.58 vs. 44.35), greater positive skewness (0.93 vs. -0.24), and higher excess kurtosis (1.39 vs. -0.27) compared to the deterministic ensemble. These statistics indicate that the stochastic approach produces softer, more asymmetric, and heavier-tailed predictions—consistent with reduced overconfidence and enhanced exposure of regional uncertainty (Figure 1). The broader distribution of logits provides a principled foundation for per-component confidence estimation (Figure 2), facilitating tunable thresholding strategies in downstream segmentation.

Conclusion

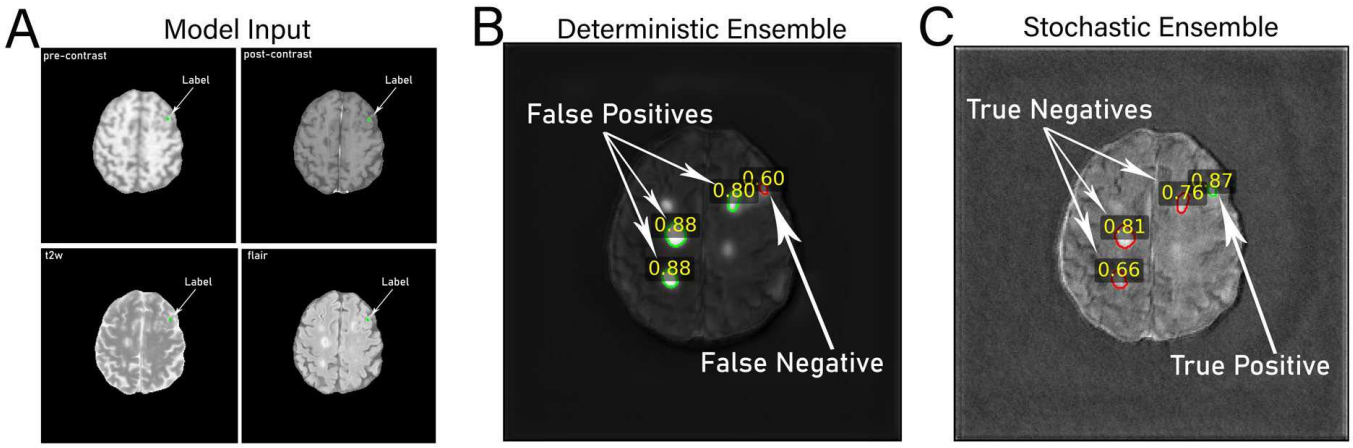
Emulating inter- and intraobserver variability in deep learning models enhances the probabilistic interpretability of model ensembles, leading to a more robust, uncertainty-aware prediction.

Statement of Impact

This work aligns model predictions with clinical uncertainty for more interpretable segmentation.



Stochastic Ensembles Reveal a Heavy-Tailed Distribution of Positive Predictions. A) Logit maps and corresponding histograms from three deterministic models (model_N: model_32, model_64, and model_256), each trained on 3D patches of size (z=16, y=N, x=N). These models produce sharply bimodal distributions, reflecting overconfident predictions in both negative and positive regions. B) Logit maps and histograms from the stochastic ensemble, obtained by averaging predictions over ten Gaussian noise–perturbed inputs per model. The resulting distribution reveals a heavy tail of intermediate logit values in the positive range, offering a probabilistic interpretation of per-blob confidence values.



Stochastic Ensembles Enable Estimation of Per-Blob Confidence. A) Input MRI sequences: pre- and post-contrast T1-weighted (T1W), T2-weighted (T2W), and FLAIR. Ground truth tumor annotations are overlaid in green. B) Deterministic ensemble

predictions yield high-confidence values in false-positive regions. Averaging logits over a region of interest (indicated by yellow labels) fails to distinguish true tumors from noise. C) In contrast, stochastic ensembles aggregate predictions across perturbed inputs, producing a more informative distribution over the logit space. Averaging over regions yields per-blob confidence estimates that more effectively discriminate true tumor regions from false positives.

Keywords
Tumor Segmentation; Neural Networks; Ensemble Learning; Uncertainty Quantification; Stochastic Inference; BraTS 2024