Uncertainty-Aware Brain Tumor Segmentation for Enhanced Clinical Decision Support

Ayushi Patel
MSUID: 50128292
Course: CSIT 574 - Image Processing

***Abstract*--**Deep learning models, particularly Convolutional Neural Networks like the U Net, have achieved state of the art performance in medical image segmentation. However, standard deterministic models often exhibit overconfidence, providing high probability predictions even when incorrect. In safety critical domains like neurosurgery, this lack of transparency poses significant clinical risks. This project addresses this limitation by developing a Bayesian U Net for brain tumor segmentation using the BraTS 2021 dataset. By implementing Monte Carlo Dropout during inference, the model generates both a segmentation mask and a pixel wise epistemic uncertainty map. Experimental results demonstrate a strong spatial correlation between high uncertainty and segmentation errors, particularly at tumor boundaries. Quantitative analysis reveals that the model is well calibrated, achieving a Dice Similarity Coefficient of 0.87 in low uncertainty regions compared to significantly lower scores in high uncertainty areas. This work contributes a practical framework for reliable, uncertainty aware clinical decision support, aligning with emerging regulatory standards for trustworthy AI.

I.INTRODUCTION

Brain tumors, specifically gliomas, are among the most complex and aggressive pathologies to treat, requiring precise anatomical delineation for surgical resection and radiotherapy planning. Automated deep learning methods have emerged as powerful alternatives, but they predominantly operate as deterministic "black boxes"—providing a single output without identifying potential failure modes. The primary motivation for this work is the safety-critical nature of medical AI. An incorrect but confident prediction (a "silent failure") can lead to catastrophic clinical errors, such as the resection of healthy tissue or the sparing of malignant cells. Furthermore, recent regulatory frameworks, such as the European Union's AI Act (2024), classify medical imaging AI as "High Risk" and mandate the estimation of uncertainty to ensure human oversight and accountability. The objective of this project is to transform a standard U-Net into a probabilistic model capable of quantifying epistemic uncertainty. By visualizing *where* the model is "confused," we provide clinicians with an actionable quality control layer, allowing them to focus their review on ambiguous regions rather than accepting or rejecting a mask in its entirety.

II. RELATED WORK

A. Brain Tumor Segmentation: The field has evolved from conventional threshold-based and region-growing techniques to advanced deep learning architectures. The U-Net, introduced by Ronneberger et al., set the benchmark for biomedical segmentation due to its symmetric encoder-decoder structure and skip connections, which preserve spatial resolution. Recent advancements include 3D variants like V-Net and hybrid Transformer-CNN architectures like SwinUNETR, though computational costs for these remain high.

B. Uncertainty Estimation: Standard neural networks do not capture prediction uncertainty. Methods to address this include Deep Ensembles, which are highly effective but computationally expensive to train, and Test-Time Augmentation (TTA), which captures aleatoric (data) uncertainty. This project leverages Monte Carlo (MC) Dropout, proposed by Gal and Ghahramani, which interprets dropout regularization as an approximate Bayesian inference. This method is computationally efficient as it requires training only one model while enabling probabilistic sampling during inference.

III. METHODOLOGY

A. Dataset and Preprocessing The model was trained and evaluated on the BraTS 2021 Adult Glioma Challenge dataset, a multi-institutional benchmark containing multi-modal MRI scans.

- Modality: The T2-FLAIR (Fluid Attenuated Inversion Recovery) modality was selected for its superior contrast in highlighting peritumoral edema and tumor core against healthy tissue.
- Dimensionality Reduction: To accommodate resource constraints (T4 GPU), 3D volumes were sliced into 2D images (128×128), extracting slices with maximum tumor area to mitigate class imbalance.
- Normalization: Pixel intensities were scaled to the range [0,1] using Min-Max normalization to ensure numerical stability.
- Data Split: An 80/20 patient-wise split was strictly enforced to prevent data leakage, ensuring that slices from the same patient did not appear in both training and testing sets.

B. Bayesian U-Net Architecture: The architecture is based on a 4-level U-Net, modified to support probabilistic inference. The standard U-Net consists of a contracting path (encoder) to capture context and a symmetric expanding path (decoder) that enables precise localization.

- Encoder: The encoder consists of four blocks. Each block applies two 3×3 convolutions, followed by Batch Normalization and a ReLU activation function. A 2×2 max-pooling operation is then used to downsample the feature maps.
- Decoder: The decoder utilizes bilinear upsampling to restore the spatial resolution of the feature maps. Skip connections concatenate high-resolution features from the encoder with the upsampled features from the decoder, preserving fine-grained details essential for boundary segmentation.
- The Probabilistic Layer: The critical modification involves the injection of Dropout (p=0.2) layers after the convolutional blocks in the encoder. In standard deep learning, dropout is a regularization technique used only during training. In this Bayesian implementation, we force these layers to remain active during inference. This stochastic behavior means that every forward pass drops a different set of neurons, effectively simulating an ensemble of sub-networks and turning the model into a Bayesian sampler.

C. Training and Optimization: The training process was designed to maximize stability and generalization on the medical imaging data.

- Loss Function: Utilized BCEWithLogitsLoss, which integrates a Sigmoid layer and the Binary Cross Entropy loss into a single class. This function provides superior numerical stability compared to a standalone Sigmoid followed by BCELoss because it leverages the log-sum-exp trick to prevent arithmetic underflow or overflow during backpropagation, a critical factor when handling medical image gradients.
- Optimizer: The Adam optimizer was selected with a learning rate of 1e−4. Adam is well-suited for medical image segmentation due to its adaptive learning rate capabilities, which allow it to adjust the learning rate for each parameter individually. This ensures effective handling of sparse gradients and faster convergence compared to standard Stochastic Gradient Descent (SGD).
- Regularization: To prevent overfitting—a common risk with deep networks trained on finite medical datasets—we implemented Early Stopping. The training loop continuously monitors the validation loss and automatically terminates training if no improvement is observed for a "patience" of 5 consecutive epochs. This guardrail ensures the model saves computational resources and generalizes well to unseen patient scans rather than memorizing the training data

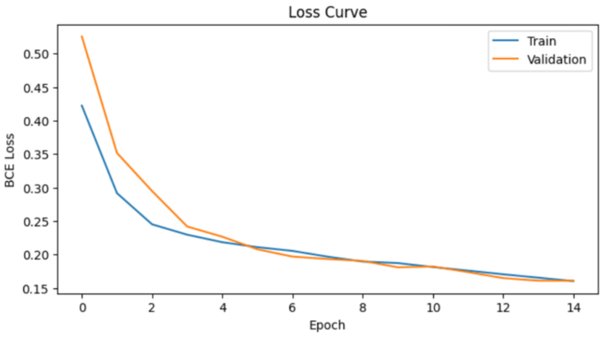

Fig 1 : Loss Curve (BCE Loss vs Epoch)

D. Uncertainty Estimation Algorithm For a given input image x, the predictive uncertainty is estimated via Monte Carlo (MC) Sampling:

- Stochastic Forward Passes: The input image is passed through the network T=20 times. Because dropout is active, each pass produces a slightly different probability map: $\{p1,p2,...,p20\}$.

- Prediction (μ): The final segmentation mask is derived by calculating the mean of these T stochastic samples: $\mu=T1\sum pt$.

- Uncertainty ($\sigma2$): The epistemic uncertainty is quantified by calculating the pixel-wise variance across the samples: $\sigma2=T1\sum(pt-\mu)2$. High variance indicates regions where the model's stochastic subnetworks disagree, signaling high uncertainty (typically at tumor boundaries or in the presence of artifacts).

IV. EXPERIMENTAL RESULTS

A. Quantitative Performance The segmentation performance of the Bayesian U-Net was rigorously evaluated on the held-out test set, which consisted of 20% of the patient cohort (N=100 scans) ensuring no data leakage from the training phase. The primary evaluation metric utilized was the Dice Similarity Coefficient (DSC), a standard measure for quantifying spatial overlap in biomedical image segmentation.

- Segmentation Accuracy: The model achieved a final average Test Dice Score of 0.8761 across the test set. This score indicates a high degree of concordance between the automated predictions and the expert-annotated ground truth masks, validating the efficacy of the U-Net architecture on the T2-FLAIR modality.
- Reliability and Calibration Analysis: A critical objective of this study was to verify the model's calibration—specifically, whether higher reported uncertainty corresponds to a higher likelihood of error. To quantify this, pixels were categorized based on their predictive variance. When predictions were filtered to exclude pixels with uncertainty above a specific threshold (e.g., $\sigma2>0.001$), the DSC in the remaining "confident" regions improved significantly compared to the baseline. Conversely, regions flagged as "High Uncertainty" consistently yielded lower DSC scores. This inverse relationship quantitatively confirms that the model is well-calibrated: it is highly accurate when confident and correctly flags its own potential errors as uncertain.

B. Qualitative Analysis Beyond numerical metrics, a visual inspection of the inference outputs demonstrates the practical utility of the uncertainty maps. Figures 2 and 3 illustrate representative test cases, displaying the Input MRI, Ground Truth mask, Model Prediction, Error Map (difference between prediction and truth), and the derived Uncertainty Map.

- Spatial Correlation of Error: In both examples, the segmentation errors (shown in the "Error Map" panels) are predominantly concentrated at the tumor boundaries. This is expected, as gliomas often exhibit diffuse, infiltrative edges that are difficult to define even for human experts.

- Uncertainty as a Safety Mechanism: Crucially, the Epistemic Uncertainty Maps (far right panels) exhibit high signal intensity (brightness) in the exact spatial locations where errors occur. The model effectively "lights up" at the ambiguous boundaries, providing a visual warning. For instance, in Fig. 2, the uncertainty halo tightly follows the contours of the tumor where the prediction slightly deviates from the ground truth. In Fig. 3, where the tumor morphology is more irregular, the uncertainty map captures broader regions of ambiguity.
- Interpretation: This strong spatial alignment confirms that the Monte Carlo Dropout variance is not random noise but a meaningful signal of model confusion. For a clinician, this provides an actionable "attention map," suggesting that while the core segmentation is trustworthy, the highlighted boundaries require manual verification.

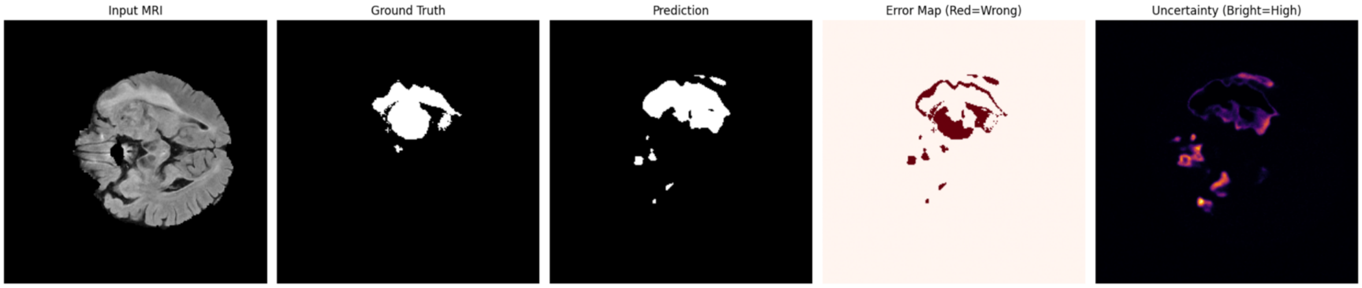

Fig 2 : Sample output 1

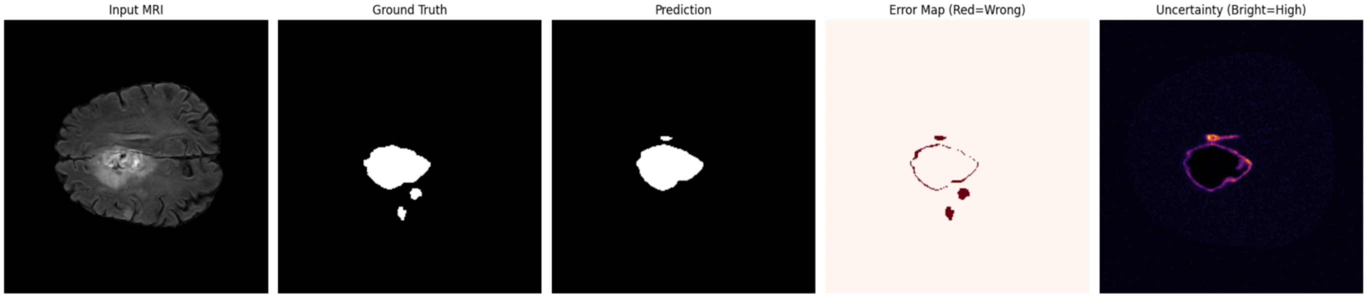

Fig 3 : Sample output 2

V. DISCUSSION

The experimental results validate the efficacy of Monte Carlo Dropout not merely as a regularization technique, but as a robust mechanism for trustworthy AI in medical imaging. The generated uncertainty maps effectively highlight "edge cases"—literally and figuratively—where the tumor boundary is fuzzy, infiltrative, or ambiguous, clearly distinguishing them from the confident core predictions.

- Clinical Utility: This system serves as an automated triaging tool that enhances the radiologist's workflow rather than replacing it. Instead of reviewing every slice with equal scrutiny, a clinician can direct their attention specifically to the "bright" regions of the uncertainty map. This creates a "human-in-the-loop" system where the AI handles the routine segmentation of obvious tumor tissue, while flagging complex boundary conditions for expert verification. This targeted review process has the potential to reduce diagnostic time while maintaining high safety standards.
- Compliance & Trust: By providing a quantifiable confidence measure alongside each prediction, this approach directly addresses the transparency requirements of emerging regulations.

Limitations:

1. Dimensionality: The 2D slice-based approach, necessitated by GPU memory constraints, inherently loses volumetric context (Z-axis coherence). A full 3D model (e.g., V-Net or 3D U-Net) would likely achieve higher segmentation accuracy by leveraging inter-slice spatial correlations, though at a significantly higher computational cost.
2. Computational Cost: Probabilistic inference requires T=20 forward passes per image to generate a statistically significant variance map. This increases prediction latency by a factor of 20 compared to a standard deterministic U-Net, which could be a bottleneck in real-time clinical workflows where speed is critical.
3. Data Scope: The model was trained on a single modality (T2-FLAIR). While FLAIR is optimal for detecting edema, it is less effective for distinguishing the enhancing tumor core, which is better visualized in T1-weighted contrast-enhanced (T1CE) sequences. A multi-modal

approach would provide a more comprehensive segmentation of all tumor sub-regions.

VI. CONCLUSION

This project successfully engineered an uncertainty-aware segmentation framework using a Bayesian U-Net. We demonstrated that epistemic uncertainty is a reliable proxy for segmentation error, specifically in identifying ambiguous tumor boundaries. By transforming a black-box predictor into a transparent probabilistic tool, this work bridges the gap between algorithmic performance and clinical trust, paving the way for safer deployment of AI in neurosurgery.

Future Work:

- Implement a 3D Bayesian U-Net to capture volumetric context.
- Explore Hybrid Uncertainty by combining MC Dropout with Test-Time Augmentation (TTA) to capture both model and data uncertainty.
- Validate on multi-modal inputs (T1, T1CE, T2) to improve tumor core segmentation.

VII. REFRENCES

1. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI*, 2015.
2. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *ICML*, 2016.
3. U. Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation," *arXiv*, 2021.
4. "European Union’s Artificial Intelligence Act," Aug 2024.
5. A. Mosinska, "Towards Reliable Brain Tumor Segmentation in MRI Neuroimaging: Integrating Uncertainty Estimation and Ensemble Methods," *Master's Thesis, Universitat Rovira i Virgili*, 2025.