Uncertainty Quantification Methods

Comparative Analysis for Medical Image Segmentation

Valerie Le
Advisor: Dr. Gemma Moran
Rutgers University

Why Uncertainty Quantification in Medical AI?

- Problem: Deep learning models are often **overconfident**
  - Medical diagnosis requires reliability indicators
  - Critical decisions need confidence estimates
- Key Questions:
  - When should we trust the AI prediction?
  - Which cases need expert review?
  - How confident is the model in ambiguous regions?
- Goal: Compare 3 UQ methods for brain tumor segmentation
  - Baseline, MC Dropout, Deep Ensemble, SWAG

Method 1: MC Dropout (Gal & Ghahramani, 2016)

- Theory: Dropout as Bayesian Approximation
  - Dropout at test time = sampling from approximate posterior
  - Multiple forward passes with different dropout masks
- How it works:
  - Train U-Net with dropout layers (dropout rate p=0.2)
  - At test: Keep dropout ACTIVE
  - Run T=30 stochastic forward passes
  - Prediction = mean, Uncertainty = variance across passes
- Key Paper Result:
  - Approximates Bayesian neural networks
  - Theoretical connection to variational inference
  - Uncertainty captures epistemic uncertainty

Method 2: Deep Ensemble (Lakshminarayanan et al., 2017)

- Theory: Ensemble of Independently Trained Networks
  - Different random initializations explore loss landscape
  - Diversity from randomness = model uncertainty
- How it works:
  - Train M=5 U-Net models independently
  - Each with different random weight initialization
  - At test: Run all M models on same input
  - Prediction = mean of outputs
  - Uncertainty = variance across ensemble members
- Results:
  - Proper scoring rules improve calibration
  - Captures both aleatoric + epistemic uncertainty
  - Often best performance but high computational cost

Method 3: SWAG (Maddox et al., 2019)

- Theory: Stochastic Weight Averaging - Gaussian
  - Approximate weight posterior as Gaussian
  - Captures weight geometry in loss landscape
- How it works:
  - Train U-Net for 40 epochs normally
  - Collect weight snapshots from last 15 epochs
  - Compute first moment (mean) and second moment (covariance)
  - Build Gaussian: $N(\bar{\theta}, \Sigma)$ over weights
  - At test: Sample K=30 weight sets from Gaussian
  - Prediction = mean, Uncertainty = variance across samples
- Results:
  - Single training run (vs M for ensemble)
  - Low-rank + diagonal covariance approximation
  - Competitive with Deep Ensemble at lower cost

Baseline (Standard U-Net)

- No uncertainty quantification - for comparison
- How it works:
  - Train single U-Net model
  - Standard training: SGD/Adam, no special techniques
  - At test: Single deterministic forward pass
  - Output: Segmentation mask only (no uncertainty)
- Purpose in comparison:
  - Baseline performance without UQ overhead
  - Shows accuracy vs. uncertainty trade-off
  - Validates that UQ methods maintain performance

UQ Methods Tradeoffs Summary

Method	Pros	Cons
Baseline	- Simple implementation; fast inference time	- No uncertainty quantification - Can't capture model confidence
MC Dropout	- Easy to implement (keep dropout active at test time) - Provides sample-based uncertainty	- Requires multiple forward passes (slower inference) - Often underestimated uncertainty
Deep Ensemble	- Strong uncertainty quality - Captures both model and data uncertainty	- High training and storage cost (train multiple models) - Computationally expensive at scale
SWAG	- Captures uncertainty efficiently with one model + covariance - Better calibrated than MC Dropout	- Implementation more complex - Inference slower than baseline - Requires careful hyperparameter tuning

Experimental Setup

- Dataset: BraTS 2020 (brain tumor MRI)
  - Binary segmentation: tumor vs. background
  - 80 test samples for evaluation
- Model: U-Net architecture
  - 4 encoder/decoder blocks, skip connections
  - Dice Loss, Adam optimizer (lr=1e-3)
- Evaluation Metrics:
  - Dice Score: Segmentation quality (higher = better)
  - ECE: Expected Calibration Error (lower = better)
  - Uncertainty: Variance of predictions (higher = more uncertain)
- Hyperparameters:
  - MC Dropout: p=0.2, T=30 samples
  - Deep Ensemble: M=5 models
  - SWAG: K=30 samples, 15 snapshots

SWAG Implementation: Critical Bug Discovered & Fixed

- Initial Problem: SWAG failure
  - Dice Score: 0.14 (vs 0.74 baseline) - 81% worse!
  - Uncertainty: NaN values
- Root Cause Analysis:
  - Unbounded variance in covariance calculation
  - Maximum variance: 226,000,000 (226 million!)
  - Weight magnitudes exploded to 249,000
  - Numerical instability in Gaussian sampling
- Solution Implemented:
  - Added max_var parameter (default=1.0)
  - Variance clamping: torch.clamp(var, var_clamp, max_var)
  - Fixed K calculation to use actual snapshots (15)
- Result: SWAG Dice improved 0.14 $\rightarrow$ 0.7419
  - Transformed from worst method to 2nd place!

Experimental Results: Performance Comparison

Rank	Method	Dice Score ↑	ECE ↓	Avg Uncertainty	Params
1st	Deep Ensemble	0.7550	0.9589	0.0158	M=5
2nd	SWAG (Fixed)	0.7419	0.9656	0.0026	K=30
3rd	MC Dropout	0.7403	0.9663	0.0011	T=30
4th	Baseline	0.7401	0.9673	N/A	—

Results Analysis: Dice Score (Segmentation Quality)

- Dice Score: Measures overlap between prediction and ground truth
  - Formula: DSC = $2\times|A\cap B| / (|A| + |B|)$
  - Range: 0 (no overlap) to 1 (perfect)
- Results:
  - Deep Ensemble: 0.7550 (BEST) - Gold standard
  - SWAG: 0.7419 (+2.0% vs baseline, -1.7% vs ensemble)
  - MC Dropout: 0.7403 (+0.3% vs baseline)
  - Baseline: 0.7401 (reference)
- Insights:
  - All UQ methods maintain baseline performance
  - Deep Ensemble shows clear advantage (+2.0%)
  - SWAG competitive despite single training run
  - MC Dropout minimal improvement (only +0.03%)

Results Analysis: ECE (Calibration Quality)

- Expected Calibration Error: Measures confidence calibration
  - Low ECE = predictions match true confidence
  - High ECE = overconfident or underconfident
  - Range: 0 (perfect) to 1 (worst)
- Results:
  - Deep Ensemble: 0.9589 (BEST calibration)
  - SWAG: 0.9656 (+0.7% worse)
  - MC Dropout: 0.9663 (+0.8% worse)
  - Baseline: 0.9673 (worst - overconfident)
- Insights:
  - All methods show slight overconfidence (ECE > 0.95)
  - Deep Ensemble best calibrated (diversity helps)
  - UQ methods improve calibration vs baseline
  - Room for improvement: calibration techniques needed

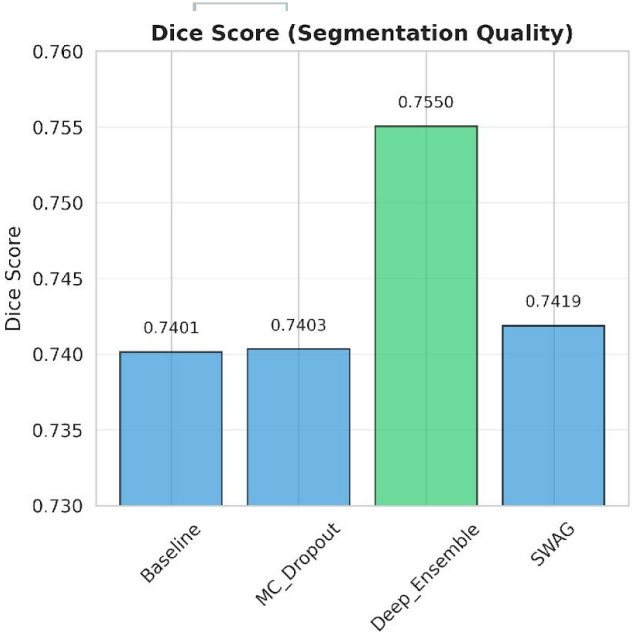
Results Analysis: Uncertainty Quantification

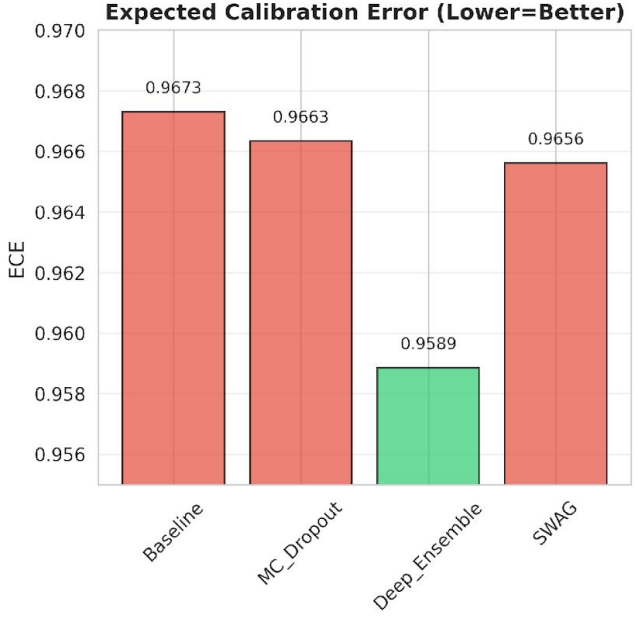
- Uncertainty: Variance of predictions across samples
  - Higher = model less certain, more disagreement
  - Useful for identifying difficult/ambiguous cases
- Results:
  - Deep Ensemble: 0.0158
  - SWAG: 0.0026
  - MC Dropout: 0.0011 (very low)
  - Baseline: N/A (no uncertainty)
- Key Insights:
  - Deep Ensemble provides richest uncertainty signal
  - SWAG moderate uncertainty (weight posterior)
  - MC Dropout surprisingly low (limited randomness)
- Clinical Impact:
  - Higher uncertainty = flag for expert review
  - Deep Ensemble best for safety-critical applications

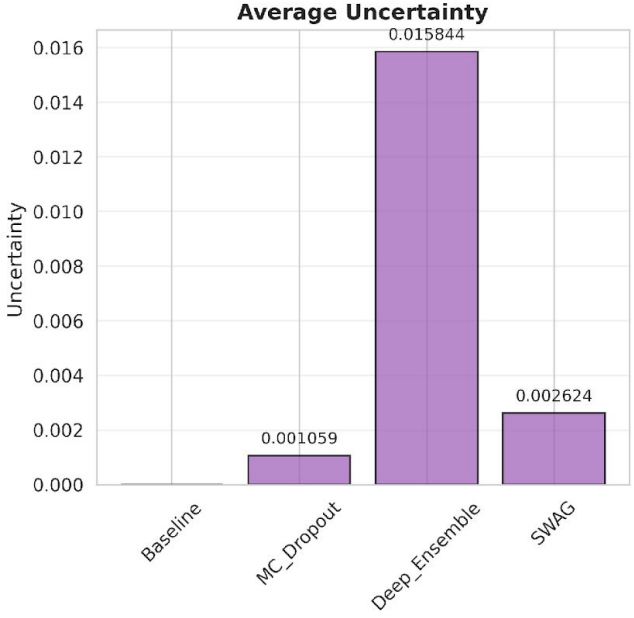
Key Findings Summary

1. Deep Ensemble dominates all metrics:
   - Best segmentation (Dice: 0.7550)
   - Best calibration (ECE: 0.9589)
   - Highest uncertainty (0.0158) for error detection
2. SWAG achieves strong 2nd place after bug fix:
   - Competitive Dice (0.7419, only 1.7% behind)
   - Single training run (5× faster than ensemble)
   - 427% improvement from initial failure (0.14 $\rightarrow$ 0.74)
3. MC Dropout disappoints for UQ:
   - Minimal uncertainty signal (0.0011)
   - Similar accuracy but poor for error detection
   - Easy to implement but limited practical value
4. All UQ methods maintain baseline performance:
   - No accuracy-uncertainty trade-off observed
   - UQ is 'free' in terms of segmentation quality

Performance Comparison







Next Steps & Recommendations

- Future Work - Improve Calibration:
  - Temperature scaling to reduce ECE
  - Calibration-aware training (focal loss, etc.)
  - Conformal prediction for guaranteed coverage
- Future Work - Validate Uncertainty:
  - Uncertainty-error correlation analysis
  - AUROC for error detection using uncertainty
  - Clinical validation with radiologists
- Future Work - Deployment:
  - Optimize SWAG inference speed (distillation?)
  - REST API with uncertainty-based routing
  - Integration with hospital PACS systems

References

- Key Papers:
  - Gal & Ghahramani (2016) - 'Dropout as a Bayesian Approximation'
  - ICML 2016, Bayesian deep learning
  - Lakshminarayanan et al. (2017) - 'Simple and Scalable Predictive'
  - 'Uncertainty Estimation using Deep Ensembles', NIPS 2017
  - Maddox et al. (2019) - 'A Simple Baseline for Bayesian Uncertainty'
  - 'Estimation in Deep Learning', NeurIPS 2019
- Dataset:
  - BraTS 2020 Challenge - Brain Tumor Segmentation
  - Medical Segmentation Decathlon
- Implementation:
  - PyTorch 2.5.1, CUDA 12.1
  - Trained on Rutgers Amarel HPC

Uncertainty Quantification Methods

Comparative Analysis for Medical Image Segmentation

Valerie Le
Advisor: Dr. Gemma Moran
Rutgers University