This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary material Reliability analysis for BraTS-GoAT segmentation: a controlled robustness study of deep-ensemble uncertainty

Riya Deepak Shet[1] and Le Zhang[2]

[1] Department of Cancer and Genomic Sciences, School of Medical Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom
[2] Department of Electronic, Electrical and Systems Engineering, School of Engineering, College of Engineering and Physical Sciences, University of Birmingham, Birmingham, United Kingdom

These tables give per-metric detail for the reliability analysis. Metric directions are constant throughout. Dice, normalized surface Dice (NSD), lesion-wise F_1 and AUROC are higher-is-better. HD95 (in millimetres), expected calibration error (ECE) and AURC are lower-is-better. WT, TC and ET are whole tumour, tumour core and enhancing tumour.

The *cross-validated single model* pools one held-out prediction for each of the 1351 training cases. The *split-0 single model* is the fold-0 network alone on that fold's 271 held-out cases. It is the like-for-like comparator for the 3-seed ensemble. Cases whose stored probability map fails a geometry-alignment check are dropped from calibration and detection, though they still contribute to Dice, HD95 and NSD. No fold excluded a case on this basis.

S1 Validation performance in full

Table S1 gives the complete result on the official BraTS-GoAT validation set for the 5-fold challenge ensemble, the configuration packaged in the submitted container. Intervals are percentile bootstrap over cases, 10 000 resamples. Denominators differ by region because the official pipeline leaves a region unscored where it is undefined in the reference.

The distance between mean and median is the substance of this table. Median Dice is above 0.89 in every region, while the mean sits 0.06 to 0.12 lower. Median HD95 is 1.4 to 2.2 mm against means of 17 to 40 mm. A small tail of catastrophic cases sets the aggregate, rather than uniform degradation. The 95% intervals on the HD95 means are correspondingly wide.

S2 Reliability across the five cross-validation folds

Table S2 gives all five folds individually, so the spread can be judged directly. Each fold is evaluated on the cases held out from its own training, and the five held-out sets partition the 1351 training cases exactly.

Table S1. Official validation-set performance of the 5-fold challenge ensemble, per case. Dice, NSD and lesion-wise F_1 are higher-is-better on a 0–1 scale; HD95 is a boundary error in millimetres, lower is better. n is the number of cases the official pipeline scored for that region.

Region	Metric	n	Mean	Median	95% CI (mean)
WT	Dice	450	0.8747	0.9368	[0.8556, 0.8923]
	NSD	450	0.4840	0.4620	[0.4638, 0.5038]
	Lesion-wise F_1	450	0.6889	0.6667	[0.6601, 0.7173]
	HD95 (mm)	450	17.12	2.24	[11.50, 23.59]
TC	Dice	449	0.8140	0.9355	[0.7888, 0.8381]
	NSD	449	0.4974	0.5243	[0.4706, 0.5234]
	Lesion-wise F_1	449	0.7379	1.0000	[0.7079, 0.7679]
	HD95 (mm)	449	25.05	2.00	[17.90, 32.77]
ET	Dice	435	0.7815	0.8970	[0.7535, 0.8086]
	NSD	435	0.5468	0.5863	[0.5214, 0.5726]
	Lesion-wise F_1	435	0.6669	0.6667	[0.6325, 0.7006]
	HD95 (mm)	435	39.97	1.41	[30.29, 50.30]

Table S2. Per-fold results for the cross-validated single model, one held-out prediction per case. Fold sizes are 271, 270, 270, 270, 270, summing to 1351. Mean and standard deviation are taken across the five folds.

Metric	Region	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Mean ± SD
Dice	WT	0.9273	0.9123	0.9330	0.9206	0.9308	0.9248 ± 0.0084
	TC	0.9132	0.8964	0.9248	0.8939	0.9182	0.9093 ± 0.0136
	ET	0.8611	0.8704	0.8771	0.8559	0.8812	0.8692 ± 0.0106
ECE	WT	0.0717	0.0732	0.0666	0.0694	0.0675	0.0697 ± 0.0028
	TC	0.0776	0.0777	0.0692	0.0827	0.0732	0.0761 ± 0.0051
	ET	0.0809	0.0769	0.0773	0.0868	0.0770	0.0798 ± 0.0043
AUROC	WT	0.8685	0.8680	0.8830	0.8717	0.8827	0.8748 ± 0.0075
	TC	0.8689	0.8695	0.8767	0.8648	0.8781	0.8716 ± 0.0056
	ET	0.8559	0.8612	0.8544	0.8509	0.8630	0.8571 ± 0.0050

The magnitudes are stable: calibration error varies by roughly 4 to 7 % of its own value across folds, and detection AUROC by under 1 %. Detection AUROC exceeds chance in 97 to 100 % of cases in every fold and region.

One qualification belongs with the table. The folds share three quarters of their training data, so this spread understates true between-fold variance. With only five folds we report a standard deviation and no confidence interval.

S3 Per-region behaviour under corruption

Table S3 gives the per-region values for the two corruptions that move segmentation quality, bias field and blur. Disagreement between ensemble members rises under bias field in every region, but not equally. The rise from the clean condition to the strongest bias is WT +26.2%, TC +28.2%, ET +16.9%.

Table S3. Per-region behaviour under the strongest level of each corruption, 271 cases. Disagreement is the variance across the three ensemble members; it has no absolute scale, so what matters is its rise relative to the clean condition.

		Single model		3-seed ensemble	
Region	Condition	Dice	ECE	Dice	Disagreement
WT	clean	0.9273	0.0716	0.9301	0.00532
	bias L3	0.9029	0.0997	0.9038	0.00672
	blur L3	0.9205	0.0803	0.9218	0.00633
TC	clean	0.9132	0.0777	0.9167	0.00562
	bias L3	0.8989	0.0900	0.9052	0.00720
	blur L3	0.8981	0.0923	0.9030	0.00827
ET	clean	0.8611	0.0808	0.8676	0.00619
	bias L3	0.8431	0.0985	0.8452	0.00723
	blur L3	0.8405	0.1033	0.8428	0.00778

Enhancing tumour, the weakest region for segmentation, is where disagreement responds least. That is a limit of the signal rather than a point in its favour: per-voxel localisation weakens where the problem is hardest.

S4 Per-case behaviour and failure modes

Per case the picture is encouraging in aggregate and less so at the tail. Detection AUROC from ensemble disagreement averages 0.871 across the 271 held-out cases, and exceeds chance in every one. Of the 20 cases where region-averaged Dice falls below 0.70, only 4 reach a detection AUROC of 0.80 or above. Across all cases the correlation between Dice and detection AUROC is +0.672. Uncertainty localises errors best on cases already segmented well, worst on those that failed.

The extremes make the point. Case 00725 segments to Dice 0.000 yet yields a detection AUROC of 0.847. Disagreement therefore ranks erroneous voxels above chance even in total failure, as a deployed alarm would need. Against that, case 01811 reaches Dice 0.458 with a detection AUROC of 0.550, barely above chance. Disagreement is therefore not presented here as a solved failure detector.