



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 detection analyses

Nair Ul Islam

GL Bajaj Institute of Technology and Management
nairulislam6@gmail.com

The analyses below use frozen probability maps from the 82-case development cohort and are conditional on the selected representation. The alternative rules and classifiers were evaluated after submission; method-specific comparisons on the blind challenge data were unavailable. Values are reported to four decimal places.

1 Development-cohort detection methods

2

Nair U Islam

Table S1: Detection performance on held-out cross-validated development predictions.
AUC and balanced-accuracy intervals are grouped-bootstrap 95% confidence intervals.
Sensitivity and specificity intervals are exact 95% confidence intervals.

Method	AUC (95% CI)	Balanced accuracy (95% CI)	Sensitivity (exact 95% CI)	Specificity (exact 95% CI)
Normalized predicted lesion-volume rule	0.9416 (0.8885 to 0.9795)	0.9043 (0.8478 to 0.9574)	0.8085 (0.6674 to 0.9085)	1.0000 (0.9000 to 1.0000)
Maximum-probability rule	0.9398 (0.8863 to 0.9794)	0.8827 (0.8094 to 0.9479)	0.8511 (0.7169 to 0.9380)	0.9143 (0.7694 to 0.9820)
Top-1024 mean rule	0.9392 (0.8843 to 0.9789)	0.9043 (0.8478 to 0.9574)	0.8085 (0.6674 to 0.9085)	1.0000 (0.9000 to 1.0000)
Maximum probability plus volume logistic regression	0.9325 (0.8720 to 0.9786)	0.8295 (0.7487 to 0.9040)	0.7447 (0.5965 to 0.8606)	0.9143 (0.7694 to 0.9820)
Fixed 14-feature logistic classifier	0.9277 (0.8634 to 0.9757)	0.8900 (0.8227 to 0.9479)	0.8085 (0.6674 to 0.9085)	0.9714 (0.8508 to 0.9993)
Reduced/core-feature logistic regression	0.9258 (0.8611 to 0.9740)	0.8757 (0.8031 to 0.9403)	0.8085 (0.6674 to 0.9085)	0.9429 (0.8084 to 0.9930)
ExtraTrees	0.9258 (0.8647 to 0.9731)	0.8757 (0.8037 to 0.9388)	0.8085 (0.6674 to 0.9085)	0.9429 (0.8084 to 0.9930)
Nested-selected logistic regression	0.9252 (0.8604 to 0.9740)	0.9043 (0.8478 to 0.9574)	0.8085 (0.6674 to 0.9085)	1.0000 (0.9000 to 1.0000)
Random forest	0.9240 (0.8615 to 0.9720)	0.8900 (0.8234 to 0.9479)	0.8085 (0.6674 to 0.9085)	0.9714 (0.8508 to 0.9993)
Nested Platt sensitivity variant	0.9198 (0.8542 to 0.9706)	0.8900 (0.8227 to 0.9479)	0.8085 (0.6674 to 0.9085)	0.9714 (0.8508 to 0.9993)
Largest-component rule	0.9170 (0.8504 to 0.9699)	0.8650 (0.7903 to 0.9322)	0.7872 (0.6434 to 0.8930)	0.9429 (0.8084 to 0.9930)
Single-feature lesion-volume logistic regression	0.8802 (0.7941 to 0.9516)	0.8511 (0.7826 to 0.9149)	0.7021 (0.5511 to 0.8266)	1.0000 (0.9000 to 1.0000)
Single-feature maximum-probability logistic regression	0.8441 (0.7334 to 0.9406)	0.7751 (0.6855 to 0.8571)	0.9787 (0.8871 to 0.9995)	0.5714 (0.3935 to 0.7368)

2 Paired comparisons with the fixed logistic classifier

Table S2: Paired development-cohort comparisons with the fixed 14-feature logistic classifier. Differences are comparator minus reference. The raw permutation and Benjamini-Hochberg-adjusted p values apply to balanced accuracy.

Comparator	AUC difference (95% CI)	Balanced-accuracy difference (95% CI)	Raw grouped permutation p	BH-adjusted p
ExtraTrees	-0.0018 (-0.0344 to +0.0267)	-0.0143 (-0.0441 to +0.0000)	1.0000	1.0000
Largest-component rule	-0.0106 (-0.0338 to +0.0084)	-0.0249 (-0.0650 to +0.0000)	0.5101	1.0000
Reduced/core-feature logistic regression	-0.0018 (-0.0096 to +0.0031)	-0.0143 (-0.0441 to +0.0000)	1.0000	1.0000
Maximum probability plus volume logistic regression	+0.0049 (-0.0280 to +0.0398)	-0.0605 (-0.1298 to +0.0035)	0.1281	0.7686
Single-feature maximum-probability logistic regression	-0.0836 (-0.1851 to +0.0015)	-0.1149 (-0.2142 to -0.0156)	0.0582	0.6984
Single-feature lesion-volume logistic regression	-0.0474 (-0.1000 to -0.0097)	-0.0389 (-0.0940 to +0.0116)	0.2261	0.9044
Maximum-probability rule	+0.0122 (-0.0134 to +0.0407)	-0.0073 (-0.0588 to +0.0401)	0.7549	1.0000
Nested-selected logistic regression	-0.0024 (-0.0200 to +0.0115)	+0.0143 (+0.0000 to +0.0441)	1.0000	1.0000
Normalized predicted lesion-volume rule	+0.0140 (-0.0006 to +0.0377)	+0.0143 (+0.0000 to +0.0441)	1.0000	1.0000
Random forest	-0.0036 (-0.0386 to +0.0257)	+0.0000 (-0.0429 to +0.0429)	1.0000	1.0000
Nested Platt sensitivity variant	-0.0079 (-0.0269 to +0.0086)	+0.0000 (+0.0000 to +0.0000)	1.0000	1.0000
Top-1024 mean rule	+0.0116 (-0.0018 to +0.0319)	+0.0143 (+0.0000 to +0.0441)	1.0000	1.0000

Nonsignificant comparisons in Table S2 are not evidence of statistical equivalence.

3 Probability diagnostics

Table S3: Probability diagnostics from held-out cross-validated development predictions. The Brier-score intervals are grouped-bootstrap 95% confidence intervals.

Method	Brier score (95% CI)	Log loss	Eight-bin ECE	Calibration intercept	Calibration slope
Fixed 14-feature logistic classifier	0.1143 (0.0839 to 0.1492)	0.4487	0.1274	0.4270	0.5088
Nested Platt sensitivity variant	0.1155 (0.0781 to 0.1574)	0.4682	0.1319	0.3025	0.4565

Nested Platt scaling did not improve the reported development diagnostics.

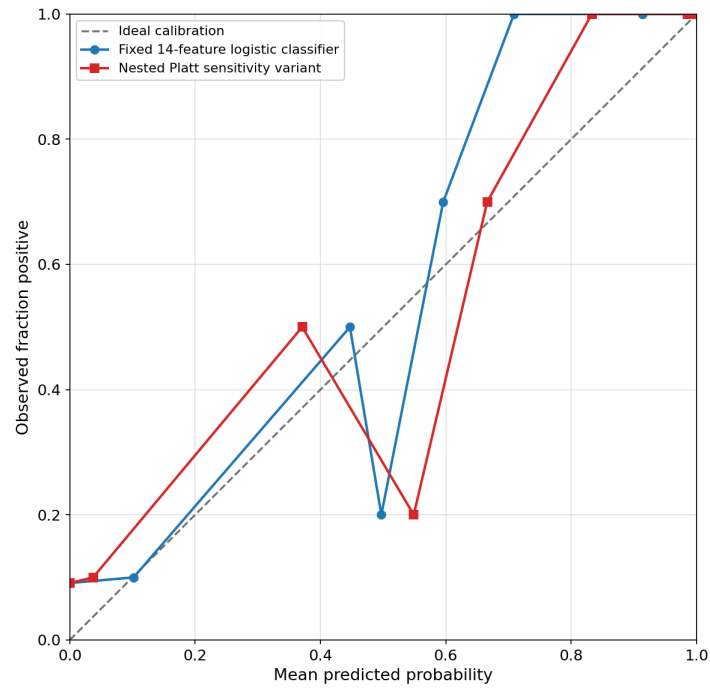


Fig. S1. Reliability of development-cohort probability outputs. Points summarize held-out cross-validated development predictions in eight equal-count bins for the fixed 14-feature logistic classifier and the nested Platt sensitivity variant. The diagonal marks perfect agreement between mean predicted probability and observed positive fraction.