

Supplementary Table S7. Comparison of the two tumor-cell detection algorithms*InstanSeg (primary) versus the legacy watershed algorithm, mean-threshold H-scores (n = 528 cores)***(A) Detected tumor cells and resulting H-scores**

Metric	InstanSeg	Legacy (watershed)	Agreement
Detected tumor cells per core, median (IQR)	1,690 (680–3,836)	1,842 (822–4,212)	-
Per-core count ratio (InstanSeg/Legacy), median			0.911
Spearman rank correlation of counts			0.947
H-score, mean $\pm$ SD	78.1 $\pm$ 78.6	86.7 $\pm$ 78.3	-
H-score, median	45.5	57.2	-

(B) Agreement of H-score values

Metric	Estimate	95% CI	Interpretation
Pearson correlation	0.996	-	Near-perfect linear association
Spearman rank correlation	0.994	-	Near-perfect monotonic association
ICC (2,1), absolute agreement	0.990	0.988–0.992	Excellent agreement
ICC (2,1), consistency	0.996	-	Excellent agreement
Lin concordance correlation	0.990	-	Excellent agreement
Bias (InstanSeg - Legacy)	-8.60	-	InstanSeg systematically lower
95% limits of agreement	-22.2 to +5.0	-	-

(C) Discrimination of the manual three-tier interpretation

Detector	Estimator	VUS	95% CI	Reference groups
InstanSeg	Normal-distribution	0.927	0.906–0.949	null 73/wild-type 319/overexpression 136
InstanSeg	Empirical (distribution-free)	0.972	0.955–0.985	Same
Legacy (watershed)	Normal-distribution	0.944	0.926–0.962	Same
Legacy (watershed)	Empirical (distribution-free)	0.967	0.948–0.983	Same

(D) Agreement of the tumor-cell classifier with the consensus ground truth

Detector	Accuracy	Cohen's κ	Reference
InstanSeg	0.904 (0.875–0.927)	0.807 (0.755–0.859)	Supplementary Table S2
Legacy (watershed)	0.824 (0.786–0.857)	0.649 (0.578–0.720)	Supplementary Table S2

CI, confidence interval; ICC, intraclass correlation coefficient; IQR, interquartile range; SD, standard deviation; VUS, volume under the ROC surface; κ , Cohen's kappa.

^aBoth algorithms were applied to the same annotated tumor regions, and H-scores were computed using the mean of the three DAB intensity-threshold calibrations. The two detection pipelines were built independently, each with its own classifier and its own threshold calibration.

^bInstanSeg detected approximately 9% fewer tumor cells per core than the legacy watershed algorithm (median per-core ratio 0.911), while the rank ordering of cores was preserved (Spearman $\rho = 0.947$).

^cThe systematic negative bias reflects the smaller number of cells detected by InstanSeg and indicates that intensity cutoffs are not directly transferable between detection algorithms.

^dVUS, volume under the receiver-operating-characteristic surface, quantifies how well the continuous H-score discriminates the three manual interpretation categories without reference to any cutoff. The VUS was estimated under both a normal-distribution and an empirical (distribution-free) estimator; because the H-score is bounded at zero and right-skewed, the empirical estimator is the more appropriate. Under either estimator, cutoff-independent discrimination was comparable between the two algorithms, with overlapping confidence intervals; in a paired case-resampling bootstrap the between-detector difference in the empirical VUS was +0.005 (95% CI, 0.000 to +0.012). 95% confidence intervals by bootstrap resampling.

^eInstanSeg was adopted as the primary detector because, as a deep-learning method, it requires no dataset-specific parameter tuning. The legacy watershed algorithm was retained as a robustness comparator (panel C).