

Supplementary Table S2. Validation of the tumor-cell classifier against a two-pathologist consensus ground truth

Cell-level validation, independent of classifier training. Cells were sampled with a fixed seed from tissue-microarray cores not used for training; two pathologists labeled each cell blinded to the classifier output, and discordant labels were adjudicated to a consensus ground truth.

(A) Overall classification performance

Detector	n used	Accuracy (95% CI)	Balanced acc.	MCC	Cohen's κ (95% CI)
InstanSeg	500	0.904 (0.875–0.927)	0.906	0.808	0.807 (0.755–0.859)
Legacy (watershed)	444 ^a	0.824 (0.786–0.857)	0.825	0.656	0.649 (0.578–0.720)

(B) Per-class performance (one-vs-rest)

Detector	Class	Support	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	F1
InstanSeg	Tumor cells	225	0.924 (0.882–0.952)	0.887 (0.844–0.919)	0.870 (0.822–0.907)	0.897
InstanSeg	Other	275	0.887 (0.844–0.919)	0.924 (0.882–0.952)	0.935 (0.898–0.959)	0.910
Legacy	Tumor cells	223	0.753 (0.693–0.805)	0.896 (0.849–0.930)	0.880 (0.826–0.918)	0.812
Legacy	Other	221	0.896 (0.849–0.930)	0.753 (0.693–0.805)	0.783 (0.728–0.829)	0.835

(C) Confusion matrices (rows = consensus ground truth, columns = classifier prediction)

InstanSeg (n = 500)

GT/Prediction	Other	Tumor cells
Other	244	31
Tumor cells	17	208

Legacy watershed (matched = 500; 56 Unclassified predictions excluded from the metrics, n = 444)

GT/Prediction	Other	Tumor cells	Unclassified
Other	198	23	49
Tumor cells	55	168	7

CI, confidence interval; GT, ground truth; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; κ , Cohen's kappa.

^aFor the legacy watershed classifier, 56 of the 500 matched cells (49 ground-truth Other, 7 ground-truth Tumor) were left Unclassified by the classifier and were excluded from the accuracy and kappa, which were therefore computed on n = 444; the InstanSeg classifier left no cell unclassified (n = 500). Macro-averaged F1 was 0.904 (InstanSeg) and 0.824 (legacy).

^bSensitivity = recall and PPV = precision; per-class metrics are one-vs-rest. Balanced accuracy is the mean of the class sensitivities. NPV values and the full set of 95% confidence intervals are provided in the validation report (Supplementary Software).

^c95% Confidence intervals for proportions by the Wilson score method; for Cohen's kappa by the normal approximation. Interobserver agreement between the two pathologists for the cell-level ground-truth labels was substantial (weighted $\kappa \approx 0.87\text{--}0.88$), so the InstanSeg classifier ($\kappa = 0.807$) approaches the reproducibility of the reference labels themselves.

^dThe InstanSeg classifier was adopted as the primary detector; the legacy watershed classifier was validated by the identical procedure and is reported here as a robustness comparator. Its lower agreement (accuracy 0.824; κ 0.649) and the 56 unclassified cells are consistent with the choice of InstanSeg as the primary detector.