The data files are available at https://github.com/fufu1209/pjp-crc-resource-aware-pathology.

File	Contents
revision1_confusion_matrices_all_models.csv	Full 9 × 9 confusion matrix for **every** model, not only TinyCNN: ColorTexture-SGD, ColorTexture-RF, TinyCNN seeds 42/7/123, ColorTexture-RF-224 and Lite224CNN-224
revision1_per_class_metrics_all_models.csv	Precision, recall, F1 and support per tissue class for every model and split
revision1_calibration_bins.csv	Per-bin observation count, mean confidence, empirical accuracy and gap, under **both** equal-width and equal-frequency binning, for every model — the reliability-diagram source data
revision1_calibration_classwise.csv	One-vs-rest class-wise ECE, Brier, mean predicted probability and observed prevalence for all nine classes
revision1_bootstrap_ci_2000.csv	2,000-replicate conditional patch-level intervals for every model and metric
revision1_paired_bootstrap_differences.csv	Paired bootstrap intervals for every model pair and metric, with an excludes_zero flag and the proportion of resamples favouring the first model
revision1_perturbation_robustness.csv	All 15 perturbations × 3 models, with accuracy, macro-F1, balanced accuracy, macro-AUROC, ECE and Brier
revision1_class_colour_signature.csv	Per-class mean and SD of each RGB channel, and dynamic-range clipping fractions, for the train and test splits

Calibration definitions

These are stated here so the numbers can be recomputed exactly.

- **Confidence** = top-label confidence, the maximum predicted class probability.
- **ECE** = $\Sigma_b (n_b / N) \cdot |acc_b - conf_b|$ over 10 bins. Reported under equal-width and equal-frequency binning; both are in the summary file.
- **MCE** = $\max_b |acc_b - conf_b|$.
- **Multiclass Brier** = mean over images of $\Sigma_k (p_{ik} - y_{ik})^2$, the sum-of-squares form, range 0–2, **not** divided by the number of classes.
- **Post-hoc calibration**: none. All probabilities are raw model outputs, and the validation split was not used to fit any calibration transformation.

Two observations worth flagging

Debris is the worst-calibrated class for both models (one-vs-rest ECE 0.035 for TinyCNN, 0.058 for ColorTexture-RF), and it is also the only class in the test split showing dynamic-range clipping (2.45% of pixels at the 8-bit ceiling; 4.7% of patches with more than 10% saturated pixels, against 0% for all eight other classes).

TinyCNN systematically under-predicts cancer-associated stroma: mean predicted probability 0.037 against an observed prevalence of 0.059, consistent with its stroma recall of 0.406.