This supplement describes patient selection, image processing, feature extraction and statistical analysis in detail. It also presents additional results and the reanalysis of the original method. References are numbered as in the main manuscript.

Supplementary Methods 1 Patient selection and image preparation

We analyzed one image from each of 944 patients: 644 from TCGA and 200 from CPTAC for development, and 100 from CPTAC for testing (Table 1). The holdout manifest was fixed on 18 October 2026. Holdout images and labels were excluded from model fitting, configuration selection and threshold selection. The selected configurations and fitted models were locked on 25 October 2026, before holdout evaluation. The later choice to focus the manuscript on the engineered model family is described in Supplementary Methods S3. The retrospective secondary analysis was conducted from February 2025 to November 2026. TCGA pathology images were downloaded in January 2025 and CPTAC pathology images in October 2026; the later GDC clinical-metadata retrieval on 11 October 2026 is distinct from the initial image download.

The 200 CPTAC development patients had previously been evaluated with the original frozen pipeline. We subsequently used these patients for supervised adaptation and tested the adapted model in separate patients from the same CPTAC study. The split did not separate recruitment institutions or scanners. It therefore assesses performance in new patients from CPTAC, not independent external validation. Shared source-specific acquisition, staining or other study characteristics may affect performance; independent multicentre validation is required before clinical application.

TCGA cohort and retrospective case linkage

The TCGA cohort comprised 323 LUAD cases (class 1) and 321 LUSC cases (class 0). The image files had been renamed with integers in place of the original slide identifiers. We recovered the corresponding case identities by subtracting 1000 from LUAD filenames and 2000 from LUSC filenames, then matching each number to its sequential entry in the retained clinical workbook. This numbering rule was confirmed during revision and yielded 644 unique TCGA case IDs and GDC case UUIDs. Public project assignments agreed with the histology labels. The exact slide/file identifiers originally selected could not be recovered.

The retained workbook contained 400 entries for each histology. Of these, 77 LUAD and 79 LUSC cases were excluded because manual pen or brush markings obscured the image or the image clarity was inadequate. The reason for each individual exclusion and the number excluded for each reason were unavailable. We obtained age, sex at birth, smoking status, pathological stage and tissue-source-site codes from GDC metadata retrieved on 11 October

2026 (Table S1). Records of the original physical scale, selected scan level and image conversion remain incomplete.

CPTAC eligibility and patient selection

CPTAC-LUAD and CPTAC-LSCC pathology images were obtained through The Cancer Imaging Archive and Imaging Data Commons [27,28], using the specimen metadata snapshot dated 9 July 2024 and IDC index release 25.0.1. Eligible slides required an exact match between slide and patient metadata, formalin-fixed paraffin-embedded tumor tissue, an acceptable tumor segment, a specific diagnosis compatible with adenocarcinoma or squamous cell carcinoma, and an IDC primary-neoplasm designation. We excluded adenosquamous, unresolved differential, unspecified, neuroendocrine and spindle-cell diagnoses. This left 1044 eligible slides from 402 patients, including 207 with LUAD and 195 with LUSC.

For each patient, we retained the first eligible slide in lexicographic order. Patients were then ordered within each histology by SHA256("20261018:Case_ID"). The first 100 per class formed the earlier 200-patient cohort, which was subsequently used for development. We ordered the remaining 202 eligible patients by SHA256("repair_holdout_20261018:Case_ID") and selected the first 50 per class for the 100-patient holdout. The other 102 patients were not analyzed. Patient selection used the predefined metadata criteria and histology strata, without reference to model predictions. All selected holdout patients were retained through download, conversion and scoring; none was replaced.

We checked the development and holdout manifests for overlapping patient identifiers and compared exact DICOM hashes between the two CPTAC subsets. Neither check identified overlap. GDC links were available for 56 holdout patients, and none matched a TCGA case UUID; checks for the remaining patients used public cohort and specimen identifiers. Histology labels were taken from the public specimen records without a new local diagnostic review. These checks support patient separation but do not establish independence by institution or scanner. Table 1 gives the sizes of the analyzed subsets.

Image preparation and tissue patch sampling

For each CPTAC image, we selected a stored DICOM volume level with a complete frame and a resolution at least as high as the intended output. TILED_FULL tiles were assembled in row-major order. The assembled image was resized with Lanczos interpolation to a longest edge of no more than 10000 pixels, preserving the complete-frame aspect ratio and using the level-0 geometry to determine the target shape. JPEGs were saved at quality 95 with the recorded Pillow defaults. We did not apply an ICC transformation or stain normalization. All 100 new DICOM files passed checks against the publisher's MD5 checksum and local checks

of identity and size. Ninety-nine used JPEG Baseline encoding and one used JPEG2000; decoding used the recorded Pillow/OpenJPEG runtime. The resulting images represent complete frames at stored resolution levels rather than the entire native slide pyramids.

Tissue was identified using OpenCV HSV saturation >15 and value strictly between 35 and 245 on the 0–255 scale. We estimated the tissue fraction in each candidate window from a thumbnail at approximately one-sixteenth of the source resolution. Windows were 128, 256 or 512 pixels wide, with a grid stride equal to half the window width. A window was eligible if at least 65% of its area met the tissue criterion. When fewer than eight windows qualified, we used up to 64 available candidates with the highest tissue fractions, allowing the patient to remain in the analysis. We sampled up to 64 windows per scale without replacement, using a deterministic seed derived from the SHA256 hash of the stable sample key, scale and seed 20261018. Reproducing the sampled patches requires the same image and sample key.

Selected windows were resized to 224 × 224 pixels with area interpolation for downsampling and cubic interpolation for upsampling. Each image contributed up to 192 patches. The tissue mask was not validated against pathologist annotations and may include stroma, normal tissue, inflammation, necrosis or artifacts as well as tumor. Window widths are expressed in source-image pixels; they do not imply a common calibrated scale in micrometers per pixel across the cohorts.

Supplementary Methods 2 Feature extraction

We calculated 123 descriptors per patch. At each source-image scale, each descriptor was summarized across patches by its median, population standard deviation and 90th percentile. This produced 123 × 3 summaries × 3 scales = 1107 values per image. Table 2 lists the feature families and counts. The ordered feature names and extraction code are retained with the analysis files. We re-extracted features from all development and holdout images with the same code.

Grayscale intensity was converted to floating point as $g = I/255$ before calculating g^2, avoiding unsigned 8-bit overflow. For grayscale, squared grayscale, saturation, hematoxylin optical density, eosin optical density and gradient magnitude, we calculated the mean, population standard deviation, interquartile range and percentiles 10, 25, 50, 75 and 90. Percentiles were obtained by linear interpolation. Hematoxylin and eosin values were derived with the fixed rgb2hed transform [32], without fitting a stain reference or applying stain normalization. Statistics were calculated within the patch tissue mask. If the mask contained fewer than 32 pixels, we used the whole patch.

GLCMs were calculated over each complete selected patch after quantizing grayscale into 32 levels with floor(I/8). We used symmetric, normalized matrices at distances {1,3} pixels and angles {0, π/4, π/2, 3π/4}. Contrast, dissimilarity, homogeneity, energy and correlation were averaged over the four angles separately at each distance, giving ten values. Energy is the square root of the sum of squared matrix probabilities, and homogeneity uses $1 + (i - j)^2$ in the denominator. Uniform local binary patterns were calculated with (radius, points) pairs of (1,8) and (3,24). Their histograms contained 10 and 26 bins, respectively, and were normalized over the tissue mask.

Gradient magnitude was calculated from 3 × 3 Sobel derivatives of floating-point grayscale. For the Fourier features, we centered the tissue-masked grayscale on its masked mean and applied a two-dimensional Hann window. We summed the shifted squared Fourier magnitude within radial bands (0,0.05], (0.05,0.15], (0.15,0.30] and (0.30,0.71], then divided each sum by the total nonzero-frequency power plus 10^{-12}. Radii were normalized by patch dimensions. The resulting power fractions are dimensionless image descriptors and have not been calibrated as photon-scattering coefficients. The extractor used in this study does not use the earlier mean real Fourier coefficient, which equals the top-left image pixel.

Dark components were identified by applying an Otsu threshold to grayscale values within the tissue mask. Pixels below the threshold were intersected with the mask and grouped with 8-connectivity. We excluded components that contained fewer than 4 or more than 1000 pixels, touched a patch border, or had nonpositive contour area or perimeter. Circularity was calculated as $\min(1, 4\pi A/P^2)$, where A is contour area and P is the closed contour perimeter. The bounding-box axis ratio was min(width,height)/max(width,height), and component area was the number of connected pixels. Density was the number of retained components divided by the number of mask pixels. Eight summary statistics for each of area, circularity and axis ratio, together with density, gave 25 values. If no components remained, the summary values were set to zero. These components have not been validated as nuclei.

For a sorted sample $x_0,\ldots,x_{n-1}$ and percentile p, we define $r=(p/100)(n-1)$, $k=\text{floor}(r)$ and $f=r-k$. Linear interpolation gives $q_p=(1-f)x_k+fx_{k+1}$. At an integer rank, the value is x_r; at the last endpoint, it is x_{n-1}. Population standard deviation uses denominator n. The eight summary statistics are the mean, population SD, IQR and percentiles 10, 25, 50, 75 and 90. Summaries of empty component lists were set to zero, and tissue statistics followed the fallback rule described above.

Let $P(i,j|d,\theta)$ denote a symmetric, normalized GLCM, with marginal means μ_x and μ_γ and marginal standard deviations σ_x and σ_γ. The properties are defined as follows: contrast, $\Sigma_{ij}(i-j)^2P$; dissimilarity, $\Sigma_{ij}|i-j|P$; homogeneity, $\Sigma_{ij}P/[1+(i-j)^2]$; energy, $\sqrt{(\Sigma_{ij}P^2)}$; and correlation,

$\Sigma_{i\,j}\,(i-\mu_x)(j-\mu_\gamma)P/(\sigma_x\sigma_\gamma)$. Degenerate marginals were handled by the recorded scikit-image graycoprops implementation. Each property was averaged over the four specified angles with a factor of 1/4. The two distances were kept separate. Table 2 gives the units and number of values per patch.

Supplementary Methods 3 Model training and statistical analysis

For the adapted model, we divided the 200 CPTAC development patients into five shuffled, stratified folds using seed 20261018. Each training set contained all 644 TCGA patients and 160 CPTAC patients, with the remaining 40 CPTAC patients used for validation. Variance filtering, standardization and classifier fitting were performed within each training partition. Standardized values were clipped to the fixed interval [−5,5]. All extracted values were finite, so imputation was unnecessary. We evaluated both the full set of 1107 features and a subset of 747 texture/shape features. Each feature set was combined with logistic regression at $C \in \{0.01,0.1,1\}$ or an RBF SVM at $C \in \{1,10\}$, giving ten candidates per training regime (Table S9; Tables S3A and S3B).

We selected the configuration with the highest mean AUC across the five validation folds. When an improvement was no greater than 10^{-12}, the first candidate in the search order was retained. The selected adapted model was L2-regularized logistic regression with C = 0.01, balanced class weights, a maximum of 2500 iterations and seed 20261018. For classifier fitting, TCGA patients received a sample weight of 1 and CPTAC patients a weight of n(TCGA training)/n(CPTAC training). This equalized the total sample weight of the two cohorts before the classifier's balanced class weights were applied. Variance filtering and standardization used the pooled training observations without sample weighting. We refitted the selected model on all 844 development patients. Removing three invariant descriptors left 1104 classifier inputs from the 1107 extracted features.

The classification threshold was chosen by maximizing Youden's index (sensitivity + specificity − 1) in the selected configuration's development out-of-fold predictions. We considered only finite ROC thresholds strictly between zero and one and resolved numerical ties toward 0.5. For the adapted model, threshold selection used predictions for the 200 CPTAC development patients only. The threshold was fixed at 0.5005128617652639, with probabilities at or above this value classified as LUAD. Both the fitted model and threshold were locked before holdout evaluation, as shown in Figure 1. Because the development folds were used to select the configuration, their AUCs are not unbiased nested estimates of the selected procedure. Performance in new CPTAC patients was assessed separately using the holdout.

For the TCGA-only comparator, configuration selection used five folds with no shared TCGA source-site codes between training and validation. This selected the texture/shape feature set and logistic regression with C = 0.01. Both the model and its threshold were fitted without CPTAC labels. We also evaluated the original 30-feature SVM, which had been frozen after fitting on TCGA and retained its original feature calculations and prediction rule. All three models were tested on the same holdout patients. Since adaptation changed the feature representation, classifier and training data relative to the original SVM, any gain cannot be attributed to correction of a single numerical defect.

LUAD was the positive class. We calculated AUC, accuracy, balanced accuracy, class-specific sensitivity, precision, specificity and F1, the Brier score and raw confusion counts. Conditional percentile confidence intervals were obtained from 2000 bootstrap samples (seed 20261018), drawing patients with replacement within each class while keeping fitted models and thresholds fixed. The same patient samples were used for paired AUC differences from the original SVM. These intervals do not account for uncertainty from model refitting, model-family selection, source sites, diagnosis or image conversion. We did not recalibrate the models, adjust thresholds or exclude patients after testing. Predictive values in this balanced holdout do not estimate those expected at clinical disease prevalence.

Feature distributions were plotted from development patients only. We selected the features shown in Figures 2 and S1 by the absolute magnitude of their coefficients in the fitted adapted model. The plots describe coefficient ranks, directions and class distributions; they did not influence fitting or testing and do not establish biological mechanisms. Figure 5 shows the five development-fold AUCs used to select each configuration. The TCGA source-site folds and CPTAC patient folds differ in population, training data and partitioning and are not directly comparable; the figure is descriptive rather than a paired comparison or evidence of superiority.

The broader model comparison had designated a frozen pathology encoder as the primary model within each training regime on the basis of development AUC. We chose to focus this manuscript on the engineered family after the comparative test results were available. No configuration, fitted coefficient or threshold was changed after testing. We therefore regard this analysis as exploratory. The conditional bootstrap intervals do not account for this reporting choice.

Both SVM and logistic-regression candidates used balanced class weights. The RBF SVMs used gamma="scale" and probability estimation. Table S9 lists the candidate grid and selected settings.

Supplementary Methods 4 Reanalysis of the original method

Inspection of the original extractor identified a mask covering the entire image, squaring in unsigned 8-bit arithmetic, and components selected above the Otsu threshold with problematic perimeter calculations. For example, squaring an intensity of 244 in uint8 gives 144, whereas floating-point arithmetic gives 59536. The mean real coefficient of an unnormalized two-dimensional Fourier transform equals the top-left pixel, and fftshift leaves that mean unchanged. For bright components with zero perimeter, adding 10^{-10} to the denominator can produce circularity values near 10^{11}. These calculations therefore could not be interpreted as physical measurements of tissue.

An exploratory analysis using only five background/corner descriptors gave an AUC of 0.680 in TCGA patient cross-validation, 0.597 in source-site cross-validation and 0.452 in CPTAC development. Some associations between the original shape features and histology also reversed direction between TCGA and CPTAC development. Acquisition effects and unstable feature calculations may therefore have contributed to the original findings, although this analysis cannot quantify their individual effects. The background/corner descriptors came from a subset previously selected using all TCGA cases, which limits interpretation of their cross-validation results.

We repeated the TCGA analysis using 644 linked cases and the original matrix of 1059 features. Five shuffled, stratified outer folds (seed 20) and three inner folds per outer split (seed 100 + fold number) kept missing-value handling, standardization, ranking by a 100-tree random forest, top-30 selection and tuning within training data. Each case received one outer-test prediction. In a separate analysis, we kept 28 conservatively harmonized source-name groups from the metadata separate in both inner and outer folds. We also removed 108 outputs from the five threshold-dependent routines, leaving 951 candidate features. To examine mask sensitivity, we used 24 sampled images with two fixed-position 1024-pixel patches per image. LAB bounds and Otsu thresholds were varied by five encoded units, and H/E concentrations by ±10%, with an HED roundtrip as the control for the latter. These analyses used the recovered original extractor and do not estimate performance or validate segmentation for the 1107-feature model. The source groups have not been verified as distinct scanners or institutions, and agreement between automatic masks does not measure accuracy against expert annotations.

Supplementary Methods 5 Software and reproducibility

The 1107-feature analysis used Python 3.12.7, NumPy 1.26.4, SciPy 1.13.1, pandas 2.2.2, scikit-learn 1.6.1 [31], joblib 1.4.2, OpenCV 4.11.0, scikit-image 0.25.0 [32] and Pillow

10.4.0. Image decoding used JPEG 9e without libjpeg-turbo and OpenJPEG 2.5.2. Refitting from the archived development matrices reproduced the selected model's holdout probabilities exactly. Rerunning prediction preserved all class assignments, with only floating-point differences in probabilities. A prepared-image check of the TCGA-only engineered pipeline also reproduced its saved prediction. Reproducing the JPEG pixels requires the same decoder build as well as the package versions. Hashes of the model, extraction code, feature schema, images and manifests identify the files used in the analysis. Tables S5–S8 give the nested reanalysis and sensitivity checks for the original TCGA method.

We retained the development and holdout feature matrices, patient manifests, candidate results, fitted pipelines and fixed predictions in the local analysis archive. The scripts reproduce_training.py and reproduce_predictions.py rerun numerical fitting and prediction. A separate package contains engineered_CPTAC_adapted, the ordered feature schema, extraction code and predict.py for CPU inference from prepared images. The model's SHA256 hash is 95b5f9757c039aee88410599bfa8a7033bf4caff00c19abff2e9c3997ef88f0b. The original image key must be retained to reproduce patch selection. These files have not yet been deposited in a public repository.

Supplementary Table 1 Linked TCGA cohort characteristics

Characteristic	LUAD (n = 323)	LUSC (n = 321)
Linked patients and images	323	321
Age in years, median [IQR]	66.7 [59.6, 73.6]	68.7 [62.4, 74.1]
Age missing	17 (5.3%)	16 (5.0%)
Female	177 (54.8%)	88 (27.4%)
Male	146 (45.2%)	233 (72.6%)
Never smoker	39 (12.1%)	12 (3.7%)
Former smoker	206 (63.8%)	232 (72.3%)
Current smoker	70 (21.7%)	68 (21.2%)
Smoking missing	8 (2.5%)	9 (2.8%)
Stage I	181 (56.0%)	156 (48.6%)
Stage II	80 (24.8%)	98 (30.5%)
Stage III	42 (13.0%)	61 (19.0%)
Stage IV	17 (5.3%)	4 (1.2%)
Stage missing	3 (0.9%)	2 (0.6%)
Tissue-source-site codes, n	32	32

Values are n (%) of the 323 LUAD or 321 LUSC cases unless stated otherwise. Age is reported as median [IQR] in years. Age at diagnosis was obtained from the primary diagnosis and converted from days using 365.25 days/year; quartiles were calculated by linear interpolation. Sex at birth and smoking status are GDC fields. Former smoking combines the three reformed-smoker categories, while missing smoking data remain separate from never smoking. Stage groups combine the reported substages, with missing stage shown separately. No primary-diagnosis values conflicted. Tissue-source-site counts refer to metadata codes and do not represent verified independent institutions. Of 400 entries per histology, 77 LUAD and 79 LUSC cases were excluded because of obstructive pen/brush markings or inadequate image clarity. Case-specific reasons and counts by reason were unavailable. Data were retrieved from the GDC cases API on 11 October 2026 and linked through the retained TCGA workbook using the numbering rule confirmed during revision. Scanner details, spatial calibration and exact selected-slide identifiers remain unavailable.

Supplementary Table 2 Candidate models and selected adapted settings

Item	Specification
Feature banks	All 1107 descriptors; 747 texture/shape descriptors
Logistic regression grid	L2; C = 0.01, 0.1 or 1; balanced class weights; max_iter = 2500
SVM grid	RBF kernel; C = 1 or 10; gamma = scale; balanced class weights; probability = True
Candidate count	10 per regime; 20 total across TCGA-only and adapted regimes
Selection	Mean five-fold validation AUC; first enumerated candidate for ties within 10^{-12}
Training transforms	VarianceThreshold; StandardScaler; fixed clipping to [−5,5]
Adapted sample weights	Classifier sample weights: TCGA 1; CPTAC n(TCGA train)/n(CPTAC train), multiplied by balanced class weights. Variance filtering/scaling are unweighted.
Selected adapted model	All descriptors; logistic regression C = 0.01; 1104 variable inputs in final fit
Threshold	0.5005128617652639 from the selected configuration's 200 CPTAC development OOF scores; LUAD if probability ≥ threshold
Seeds	20261018 for folds and estimators; stable SHA256 seed for patch sampling

Supplementary Tables 3 Nested TCGA evaluation of the original feature set

Model	Mean-fold AUC ± SD	AUC mean CI	LUAD recall	LUSC recall
SVM	0.840 ± 0.034	0.810–0.870	0.718	0.850
Logistic regression	0.835 ± 0.034	0.805–0.867	0.734	0.816
Decision tree	0.772 ± 0.033	0.735–0.808	0.731	0.704
Gradient boosting	0.840 ± 0.027	0.811–0.871	0.746	0.782

This analysis used the original 1059-feature set and 644 TCGA patients, with top-30 feature selection and tuning restricted to training data within nested cross-validation. Recall was calculated from pooled out-of-fold predictions. Confidence intervals for mean AUC used 2000 patient resamples within fold and class, conditional on the fitted predictions. These results concern the original method and do not estimate internal performance of the adapted model.

Supplementary Tables 4 Nested evaluation by source group

Model	Mean-fold AUC ± SD	Pooled AUC [95% CI]	LUAD recall	LUSC recall
SVM	0.808 ± 0.086	0.791 [0.720, 0.868]	0.684	0.804
Logistic regression	0.805 ± 0.080	0.791 [0.724, 0.868]	0.700	0.776
Decision tree	0.708 ± 0.075	0.714 [0.664, 0.760]	0.582	0.692
Gradient boosting	0.785 ± 0.087	0.774 [0.708, 0.843]	0.684	0.729

The original pipeline was evaluated using 28 harmonized source groups from the metadata, with no group shared between training and validation/test partitions in either the inner or outer folds. Pooled-AUC intervals were calculated by resampling source groups 1000 times while keeping predictions fixed. These groups have not been verified as distinct scanners or independent institutions.

Supplementary Tables 5 Effect of removing threshold-dependent features

Model	Mean-fold AUC ± SD	Pooled AUC	AUC difference [95% CI]	LUAD recall	LUSC recall
SVM	0.808 ± 0.034	0.808	-0.031 [-0.055, -0.012]	0.703	0.832
Logistic regression	0.812 ± 0.031	0.810	-0.025 [-0.048, -0.005]	0.706	0.810
Decision tree	0.740 ± 0.022	0.742	-0.026 [-0.071, 0.022]	0.650	0.748
Gradient boosting	0.820 ± 0.045	0.816	-0.013 [-0.033, 0.007]	0.759	0.779

Removing 108 outputs left 951 candidate descriptors from the original feature set. We repeated nested training using the original patient folds. AUC differences were calculated as pooled AUC after removal minus the original pooled AUC. Intervals used 1000 paired patient resamples within class; they were unadjusted and did not include model refitting. This analysis does not validate the dark-component method used in this study.

Supplementary Tables 6 Leading model features in TCGA

Exact feature identifier	Coefficient	LUSC median [Q1, Q3]	LUAD median [Q1, Q3]
s512_median_fft_power_0_0.05	-0.1555	0.191 [0.148, 0.254]	0.147 [0.112, 0.208]
s512_q90_fft_power_0_0.05	-0.1282	0.328 [0.255, 0.413]	0.243 [0.193, 0.332]
s128_median_fft_power_0.3_0.71	+0.1139	0.0455 [0.0304, 0.0587]	0.055 [0.0402, 0.074]
s512_std_lbp_r3_4	+0.1117	0.00161 [0.00135, 0.00199]	0.00141 [0.00122, 0.00179]
s256_q90_dark_component_axis_ratio_iqr	+0.1073	0.333 [0.315, 0.335]	0.333 [0.313, 0.336]
s256_std_dark_component_axis_ratio_q25	-0.1052	0.0485 [0.0395, 0.0596]	0.0416 [0.0356, 0.0503]
s512_std_gradient_iqr	+0.1046	0.0738 [0.0599, 0.0967]	0.0916 [0.0704, 0.118]
s512_median_fft_power_0.05_0.15	+0.1045	0.313 [0.262, 0.348]	0.289 [0.233, 0.336]
s512_std_fft_power_0_0.05	-0.1033	0.0857 [0.0634, 0.113]	0.0659 [0.0497, 0.0866]
s256_median_eosin_OD_q10	-0.1014	0.00423 [0.00274, 0.00678]	0.00584 [0.00329, 0.00979]
s256_median_glcm_correlation_d1	+0.0998	0.875 [0.826, 0.905]	0.827 [0.762, 0.892]
s512_median_dark_component_density	-0.0994	0.00356 [0.00209, 0.00567]	0.00373 [0.00221, 0.00622]

Supplementary Tables 7 Leading model features in CPTAC development

Exact feature identifier	Coefficient	LUSC median [Q1, Q3]	LUAD median [Q1, Q3]
s512_median_fft_power_0_0.05	-0.1555	0.224 [0.156, 0.269]	0.15 [0.111, 0.192]
s512_q90_fft_power_0_0.05	-0.1282	0.352 [0.238, 0.43]	0.238 [0.179, 0.31]
s128_median_fft_power_0.3_0.71	+0.1139	0.0443 [0.031, 0.0611]	0.0607 [0.0397, 0.0811]
s512_std_lbp_r3_4	+0.1117	0.00114 [0.000871, 0.0014]	0.00122 [0.00101, 0.00154]
s256_q90_dark_component_axis_ratio_iqr	+0.1073	0.333 [0.333, 0.35]	0.333 [0.333, 0.35]
s256_std_dark_component_axis_ratio_q25	-0.1052	0.0368 [0.0302, 0.0448]	0.0354 [0.029, 0.0413]
s512_std_gradient_iqr	+0.1046	0.0576 [0.0378, 0.0805]	0.0711 [0.052, 0.0938]
s512_median_fft_power_0.05_0.15	+0.1045	0.183 [0.161, 0.213]	0.213 [0.182, 0.26]
s512_std_fft_power_0_0.05	-0.1033	0.0863 [0.0653, 0.103]	0.0652 [0.0517, 0.0812]
s256_median_eosin_OD_q10	-0.1014	0.00826 [0.00509, 0.013]	0.00914 [0.00568, 0.0121]
s256_median_glcm_correlation_d1	+0.0998	0.698 [0.619, 0.757]	0.719 [0.673, 0.775]
s512_median_dark_component_density	-0.0994	0.0035 [0.0022, 0.00608]	0.00361 [0.00231, 0.00516]

The same 12 features are shown for both cohorts, ranked by absolute coefficient magnitude in the fitted adapted model. The prefixes s128/s256/s512 indicate source-window widths in pixels, and median/std/q90 indicate aggregation across patches. Suffixes identify the statistics calculated within each patch, as defined in Table 2. Fourier fractions, LBP frequencies, axis ratios and correlation are dimensionless. The feature eosin_OD uses fixed-transform units; gradient is expressed in normalized image-coordinate units, and component density is count/mask pixel. Coefficients apply after scaling and clipping based on the training data. These raw distributions are descriptive, exclude holdout data and do not establish causal effects of individual features.

Supplementary Tables 8 Class-specific performance of the adapted engineered model

Metric	**Estimate**	**Conditional 95% CI**
LUAD precision	0.824	0.741–0.911
LUAD sensitivity	0.840	0.740–0.940
LUAD specificity	0.820	0.700–0.920
LUAD F1	0.832	0.755–0.900
LUSC precision	0.837	0.750–0.927
LUSC sensitivity	0.820	0.700–0.920
LUSC specificity	0.840	0.740–0.940
LUSC F1	0.828	0.750–0.900

The holdout contained 100 patients, with 50 in each class. Confidence intervals were calculated from the same 2000 bootstrap samples drawn within class as for the main performance estimates, keeping predictions and the threshold fixed. In this binary classification, specificity for one class equals sensitivity for the other. Precision reflects the balanced class distribution in the holdout. The intervals exclude uncertainty from model refitting and site clustering.

Supplementary Tables 9 TCGA-only source-site validation

Features	Head	Five fold AUCs in recorded order	Mean ± SD	Pooled OOF AUC
All	LR C=0.01	0.8583, 0.6698, 0.7563, 0.8395, 0.7222	0.7692 ± 0.0793	0.7584
All	LR C=0.1	0.8583, 0.6854, 0.7815, 0.7920, 0.7135	0.7662 ± 0.0684	0.7377
All	LR C=1	0.8292, 0.6714, 0.7696, 0.7606, 0.7001	0.7462 ± 0.0620	0.7131
All	SVM C=1	0.8417, 0.5561, 0.7318, 0.8558, 0.7407	0.7452 ± 0.1199	0.7262
All	SVM C=10	0.8042, 0.5507, 0.7202, 0.7612, 0.7368	0.7146 ± 0.0970	0.6716
Texture/shape	LR C=0.01	0.8500, 0.6508, 0.7794, 0.8463, 0.7346	0.7722 ± 0.0833	0.7652
Texture/shape	LR C=0.1	0.8583, 0.6564, 0.7952, 0.7931, 0.7121	0.7630 ± 0.0790	0.7325
Texture/shape	LR C=1	0.8417, 0.6435, 0.7595, 0.7527, 0.6863	0.7367 ± 0.0758	0.6915
Texture/shape	SVM C=1	0.7583, 0.5832, 0.7560, 0.8425, 0.7532	0.7386 ± 0.0947	0.7290
Texture/shape	SVM C=10	0.7792, 0.5771, 0.7696, 0.7611, 0.7262	0.7226 ± 0.0838	0.6767

Supplementary Tables 10 CPTAC patient validation with TCGA in each fit

Features	Head	Five fold AUCs in recorded order	Mean ± SD	Pooled OOF AUC
All	LR C=0.01	0.8075, 0.8450, 0.8125, 0.8550, 0.7900	0.8220 ± 0.0271	0.8150
All	LR C=0.1	0.7650, 0.8025, 0.7400, 0.8275, 0.7975	0.7865 ± 0.0342	0.7932
All	LR C=1	0.7125, 0.7700, 0.6975, 0.8150, 0.7800	0.7550 ± 0.0489	0.7601
All	SVM C=1	0.8125, 0.8200, 0.8500, 0.8825, 0.7175	0.8165 ± 0.0619	0.8025
All	SVM C=10	0.8125, 0.8125, 0.8075, 0.8975, 0.6600	0.7980 ± 0.0858	0.7772

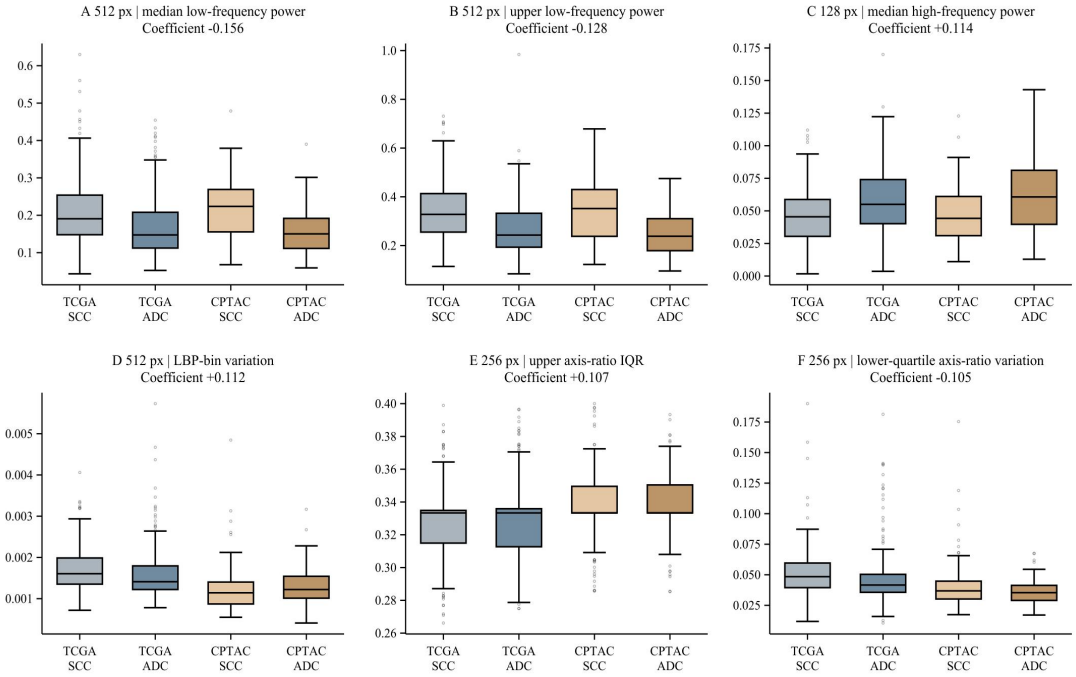
Features	Head	Five fold AUCs in recorded order	Mean ± SD	Pooled OOF AUC
Texture/shape	LR C=0.01	0.7600, 0.7725, 0.7950, 0.8350, 0.7775	0.7880 ± 0.0291	0.7808
Texture/shape	LR C=0.1	0.7500, 0.7500, 0.7500, 0.7975, 0.7500	0.7595 ± 0.0212	0.7636
Texture/shape	LR C=1	0.7125, 0.7275, 0.7050, 0.7575, 0.7200	0.7245 ± 0.0203	0.7290
Texture/shape	SVM C=1	0.8000, 0.7925, 0.7800, 0.8225, 0.6925	0.7775 ± 0.0500	0.7770
Texture/shape	SVM C=10	0.8150, 0.7425, 0.7375, 0.7925, 0.6500	0.7475 ± 0.0637	0.7490

Means are unweighted averages across five folds, and SD is the sample standard deviation (ddof=1). We selected the configuration with the highest mean AUC in each training regime. Because the same folds were used for selection, these estimates may be optimistic. The adapted model used all features with LR C=0.01; the TCGA-only model used texture/shape features with LR C=0.01. Final-holdout labels were not used to select any candidate. LR denotes logistic regression, and SVM uses an RBF kernel.

Supplementary Table11 Mask sensitivity in the original pipeline

Perturbation	Mask	Dice agreement median [IQR]	Absolute relative count change (%) median [IQR]
LAB bounds expanded 5	LAB	0.121 [0.082, 0.162]	1104.1 [577.0, 1763.7]
LAB bounds contracted 5	LAB	0.065 [0.041, 0.089]	95.6 [94.2, 97.3]
Otsu cutoff −5	Otsu	0.982 [0.970, 0.988]	8.6 [4.8, 16.1]
Otsu cutoff +5	Otsu	0.982 [0.969, 0.988]	7.6 [4.7, 19.6]
HED roundtrip	LAB	0.882 [0.625, 0.918]	12.9 [7.1, 43.7]
HED roundtrip	Otsu	0.995 [0.992, 0.998]	2.0 [1.1, 3.7]
H × 0.9	LAB	0.849 [0.732, 0.884]	26.8 [18.0, 57.4]
H × 0.9	Otsu	0.996 [0.994, 0.998]	2.2 [1.8, 3.4]
H × 1.1	LAB	0.873 [0.772, 0.910]	18.0 [14.1, 22.4]
H × 1.1	Otsu	0.996 [0.994, 0.998]	2.2 [1.4, 2.9]
E × 0.9	LAB	0.696 [0.648, 0.745]	62.4 [51.5, 85.4]
E × 0.9	Otsu	0.996 [0.994, 0.998]	3.3 [2.2, 4.2]
E × 1.1	LAB	0.748 [0.688, 0.781]	35.7 [27.2, 39.4]
E × 1.1	Otsu	0.996 [0.992, 0.998]	2.7 [1.6, 3.9]

We evaluated two patches at fixed positions in each of twenty-four sampled images. H/E perturbations were compared with the HED roundtrip reference; other perturbations were compared with the original automatic masks. Dice measures stability between automatic masks, without comparison to expert annotations. Count change was calculated as 100 × |perturbed − reference|/reference. Patches with zero reference counts were omitted when calculating within-image averages. These results apply to the original routines.



Supplementary Figure 1 Distributions of the six features with the largest absolute coefficients in the adapted model. Only development data are shown: 321 LUSC and 323 LUAD patients from TCGA and 100 per class from CPTAC. Coefficients apply to standardized and clipped inputs, with positive values favoring LUAD. Boxes span the linearly interpolated quartiles, center lines indicate medians, and whiskers extend to the most extreme values within 1.5 IQR. Points beyond the whiskers are outliers. “Upper” denotes the 90th percentile across patches and “variation” the population SD across patches. Axis-ratio IQR is calculated from components within a patch. Fourier fractions, LBP frequencies and axis ratios are dimensionless. Tables S4A and S4B give the exact feature identifiers and summaries. Holdout labels were not used in these descriptive plots.