5 Supplementary Methods

5.1 Internal-external cross-validation

Training and testing data were separated at the level of hospitals/institutions (Fig S1). To balance the size of various folds, we made sure each fold contained at least one "large" institution. Large institutions were defined as those having a minimum of 9 unique patients.

5.2 NuCLS model

Our NuCLS model modifies the Pytorch implementation of the Mask R-CNN architecture (He *et al.*, 2017).

5.2.1 Hyperparameters

We used a ResNet18 backbone that was pretrained on ImageNet. Single-GPU training was done using a batch size of 4, using a stochastic gradient descent optimiser with a learning rate of 2e-3 and a momentum of 9e-1. The learning rate and momentum were identified using grid search on the validation dataset during prototyping. All ground truth nuclei were kept per image at training, while detections were limited to a maximum of 300 nuclei at inference. 3,000 anchors were kept from the region proposal network after non-maximum suppression (NMS), using an NMS threshold of 0.7. The length of anchor sides used in pixels (relative to upsampled images, see below) is 12, 24 and 48.

5.2.2 Resize using scale factor

Mask R-CNN resizes input images to have a constant short side. While this may work for datasets where the variability in image size is modest, or where the camera distance is variable, it is not suitable in computational pathology applications where large tile sizes are favorable for efficient and scalable inference. Resizing to a constant short side would shrink nuclei during inference. To remedy this NuCLS resizes using a scale factor, instead, thus preserving the nuclear size and aspect ratio at inference for any tile size. We used a scale factor of 4.0, meaning that images were digitally zoomed to a 0.05 micron-per-pixel resolution before being analyzed. This corresponded to a sTILs diameter of 4.4 "pixels" in the feature map generated by the ResNet18 backbone. As a form of scale augmentation, we jittered this scale factor by up to 10% during training.

5.2.3 Training with hybrid datasets

Our annotation protocol generates a mixture of manually placed bounding boxes and approved suggestions of segmented nuclei. We train from this data by ignoring bounding boxes when calculating the mask loss.

5.2.4 Specialized classification convolutions

Four extra convolutional filters were applied to the feature map output from the ResNet18 backbone (He *et al.*, 2016). The filters had a kernel size of 3, a stride of 1, and a dilation and padding of 1 to preserve feature map size (Fig 4a). The resultant feature map was only used for classification and only contributed to the classification loss. The same procedure used for box regression was used for classification: 1. ROIAlign to obtain per-object convolutional feature maps; 2. flattening of the feature map; 3. passage through a single fully-connected layer.

5.2.5 Class-agnostic detection & segmentation

Both the box regression output and nucleus masks were simplified and made classification-agnostic. We relied on the fact that nucleus shapes and sizes are fairly homogeneous to simplify the learning problem and preserve classification probability vectors at inference. Specifically, we relied on a global NMS process (Fig 4b). We summed the classification probabilities for all classes (i.e. everything except background), and concatenated all these "objectness" scores for each FOV. An NMS process was then carried out as usual. That is, boxes were sorted by objectness score, and if a box overlapped with a higher-scoring box by more than a particular IOU threshold (0.2 in our case), it was removed.

5.2.6 Data augmentation

Previous research has shown that the combined use of color normalization and augmentation improves performance of deep learning models in histopathology applications (Tellez *et al.*, 2019). All FOVs were color normalized using the Macenko method before training began (Macenko *et al.*, 2009). During training, FOVs also underwent a stain augmentation routine (Tellez *et al.*, 2018). This augmentation routine randomly perturbed the hematoxylin and eosin channels each time the image was loaded, using a sigma of 0.5 for the random uniform distribution. The HistomicsTK package was used for both the color normalization and augmentation operations (digitalslidearchive.github.io). Additionally, each training image was cropped at a random location after loading to memory (300×300 pixel region) to increase robustness.

5.2.7 Handling class imbalance

Nucleus class imbalance was mitigated by weighted random sampling with replacement. With the exception of ambiguous nuclei, which received zero weight, class weights were inversely proportional to the frequency of occurrence in the training set. Since we load data on a per-FOV basis, each FOV f was assigned a sampling weight W_f that favors FOVs with a high density of uncommon nuclear classes, as follows:

$$W_f = U_f \div \sum_{i=1}^{F} U_i \tag{1}$$

$$U_f = \sum_{c=1}^{C} \left(W_c N_{cf} \right) \div A_f \tag{2}$$

Where, C is the number of classes, F is the number of FOVs in the training set, N_{cf} is the number of nuclei of class c in FOV f, and A_f is the area of FOV f. W_c is the weight assigned to class c and is determined as follows:

$$W_c = V_c \div \sum_{i=1}^{C} V_i \tag{3}$$

$$V_c = 1 \div \sum_{f=1}^{F} N_{cf} \tag{4}$$

5.2.8 Matching detections

Algorithmic detections were matched to ground truth using linear sum assignment from the Scipy library (Kuhn, 1955).

Supplementary Tables

Table S1. NuCLS model tuning for the nucleus detection task on the validation set (fold 1). All accuracy values are percentages. After passage through the model backbone, the feature map is markedly smaller than original images due to the max pooling operations. This means that without digital zooming, the diameter of a 'typical' small nucleus, say TILs, is very small in the feature map. As a consequence, when the object-specific part of the feature map is pooled using ROIAlign, there is very little information to use for box regression or classification. Abbreviations: MPP, microns-per-pixel; AP@0.5, average precision when a threshold of 0.5 is used for validating a detection.

Scale factor	**Equivalent MPP**	**Backbone**	**TILs diameter (image, pixels)**	**TILs diameter (featmap, 'pixels')**	**AP @ 0.5**
1	0.2	Resnet18	30	1.1	61.7
1	0.2	Resnet34	30	1.1	63
1	0.2	Resnet50	30	1.1	62
2.67	0.075	Resnet18	80	3	76.4
2.67	0.075	Resnet34	80	3	74.3
2.67	0.075	Resnet50	80	3	Mem.Err.
4	0.05	Resnet18	120	4.4	75
4	0.05	Resnet34	120	4.4	72.9
4	0.05	Resnet50	120	4.4	Mem.Err.

Table S2. NuCLS model tuning for the nucleus classification task on the validation set (fold 1). All accuracy values are percentages. Empty entries correspond to metrics which were not applicable for the configuration (config) being studied. Classification AUROC statistics were not possible for configs where each nucleus had a single classification as opposed to a classification probability vector, as in the baseline Mask R-CNN model. The baseline model achieves a lower performance. We show that this is due in large part to the coupling of detection and classification, which may not be ideal for datasets with many small and clustered objects. After decoupling, the performance dramatically improves. Configs where the model was trained on super-classes do not have accuracy statistics for the main classes. On the other hand, when models were trained on the main classes, super-class predictions were easily obtained by aggregating the predicted class probabilities.

Config	**Detection**	**Overall classification accuracy**						**Classification accuracy breakdown (AUROC)**								
		MCC		**Micro**		**Macro**		**Tumor**			**Stromal**			**sTILs**		
	AP @.5	Supercl.?		Supercl.?		Supercl.?		Subclasses		Superclass	Subclasses		Superclass	Subclasses		Superclass
		No	Yes	No	Yes	No	Yes	Non-mitotic	Mitotic		Stromal	Macrophage		Lymphocyte	Plasma cell	
1	70	1.8	-3	-	-	-	-	-	-	-	-	-	-	-	-	-
2	74.5	57	65	93.4	**94.3**	**85.2**	**88.2**	93.1	**91.5**	93.2	88.8	71	**83.6**	95	78.6	95
3	**75.4**	**59.6**	**66**	**93.5**	93.7	84.7	85.2	**94.2**	90.6	**94.5**	**89.1**	**73.5**	82	**95.2**	84.2	**95.7**
4	72.2	52.6	60.9	91	92.3	82.4	83.6	92.5	90.8	92.1	86.7	61.7	78.9	94.7	82.9	93.4
4+	72.2	54.5	62.5	90.3	91.9	84.1	85.8	92.2	88.5	92	88.1	68.4	81.5	93.7	**84.4**	93.4
5	72.6	-	-5	-	-	-	-	-	-	-	-	-	-	-	-	-
6	74.8	-	63.6	-	93.5	-	85.9	-	-	92.8	-	-	81.3	-	-	95
7	72.2	-	63.1	-	93.1	-	82.8	-	-	91.9	-	-	81	-	-	94.9
7+	72.2	-	64.8	-	92.7	-	83.7	-	-	93.1	-	-	83.1	-	-	94.8

Config 1: Baseline Mask R-CNN implementation. We discounted bounding boxes from the mask loss to enable training on our hybrid data.
Config 2: Config 1, but with class-agnostic detection and non-maximum suppression.
Config 3: Config 2, but with 4 extra convolutions that specialize in classification.
Config 4: Config 1 for nucleus detection, then an independent nucleus classification model using thumbnails of detected nuclei.
Config 4+: Same model from config 4, but with test-time augmentation (random shift) at the classification stage.
Config 5: Config 1 but trained using supercategories.
Config 6: Config 2 but trained using supercategories.
Config 7: Config 4 but trained using supercategories.
Config 7+: Same model from config 7, but with test-time augmentation (random shift) at the classification stage.

Table S3. Generalization accuracy of the NuCLS models trained on the corrected single-rater dataset, and evaluated on the multi-rater dataset using internal-external cross-validation. All accuracy values are percentages. Fold 1 acted as the validation set for hyperparameter tuning, so the bottom row shows mean and standard deviation of three values (folds 3-5). Note that the number of testing set nuclei varied by fold because the split happens at the level of hospitals and not nuclei. There were no testing set slides with available multi-rater truth to assess the performance on fold 2. Notice that the classification accuracy is consistently higher when the assessment was done at the level of super-classes. Abbreviations: AP@.5, average precision when a threshold of 0.5 is used for considering a detection to be true; mAP@.5:.95, mean average precision at detection thresholds between 0.5 and 0.95.

Fold	Detection			Segmentation			Classification					
	N	AP @.5	mAP @.5:.95	N	Median IOU	Median DICE	N	Super-classes?	Accuracy	MCC	AUROC (micro)	AUROC (macro)
1 (Val.)	209	62.9	21.0	42	67.6	80.7	173	No	70.5	63.6	94.2	85.6
								Yes	86.1	79.0	95.7	95.6
3	66	65.2	29.0	7	76.9	86.9	52	No	63.5	42.4	80.7	85.5
								Yes	61.5	42.5	75.1	84.7
4	317	71.5	32.6	82	76.2	86.5	278	No	68.0	54.3	94.3	89.3
								Yes	84.9	75.5	96.9	92.0
5	213	58.3	22.9	49	71.8	83.6	174	No	67.8	55.8	92.2	90.4
								Yes	75.3	65.6	91.4	95.2
Mean (Std)	-	65.0 (5.4)	28.2 (4.0)	-	74.9 (2.3)	85.7 (1.5)	-	No	66.4 (2.1)	50.8 (6.0)	89.1 (6.0)	88.4 (2.1)
								Yes	73.9 (9.6)	61.2 (13.8)	87.8 (9.2)	90.6 (4.4)

Table S4. Generalization accuracy of the trained NuCLS models - broken down by superclass. All accuracy values are percentages. Note that the corrected single-rater dataset is likely more reflective of the generalization accuracy, since it contains 1,744 unique FOVs. The multi-rater dataset only has 52 unique FOVs, hence the large variation in performance.

Fold	N	MCC				AUROC				
		Overall	Tumor	Stromal	sTILs	Micro-avg.	Macro-avg.	Tumor	Stromal	sTILs
Training: Single-rater dataset; Testing: Single-rater dataset										
1 (Val.)	5351	65.2	72.9	47.1	73.7	93.7	89.0	94.2	83.2	95.3
2	13597	68.2	73.7	53.0	76.6	94.6	86.5	94.5	87.4	96.2
3	11176	68.1	74.9	46.9	77.9	94.4	89.4	96.1	84.3	95.7
4	7288	73.5	80.6	56.9	79.6	96.1	87.4	97.2	89.1	95.9
5	6294	52.4	57.4	40.7	60.1	89.0	80.8	88.8	80.7	91.0
Mean (Std)	-	65.6 (7.9)	71.7 (8.6)	49.4 (6.1)	73.5 (7.8)	93.5 (2.7)	86.0 (3.2)	94.2 (3.2)	85.4 (3.2)	94.7 (2.1)
Training: Single-rater dataset; Testing: Multi-rater dataset										
1 (Val.)	173	79.0	88.0	73.0	78.6	95.7	95.6	97.7	94.4	95.5
3	52	42.5	38.5	26.3	73.9	75.1	84.7	87.1	83.0	90.9
4	278	75.5	77.8	53.1	90.2	96.9	92.0	96.4	91.9	99.2
5	174	65.6	60.0	67.1	72.1	91.4	95.2	96.6	92.2	97.9
Mean (Std)	-	61.2 (13.8)	58.8 (16.1)	48.8 (16.9)	78.8 (8.2)	87.8 (9.2)	90.6 (4.4)	93.4 (4.4)	89.0 (4.3)	96.0 (3.6)

Table S5. List of interpretable features used as input for DTALE, which were extracted using the HistomicsTK package.

Category	N	Description	Feature	Category	N	Description	Feature
Size	4	Pixels occupied by the nucleus	Area	**Edges**	8	Gradients and canny edge filters (hematoxylin channel)	Mag.Mean
		Length of major/minor axes of the ellipse with the same 2nd central moments	MajorAxis				Mag.Std
			MinorAxis				Mag.Skew
		Pixelated perimeter using 4-connectivity	Perimeter				Mag.Kurt.
Shape	6	Similarity to the shape of a circle	Circularity				His.Entropy
		Eccentricity of fitted ellipse (a measure of aspect ratio)	Eccentricity				His.Energy
		Diameter of a circle with the same area	Equiv.Diam.				Canny.Sum
		Ratio of nucleus area to its bounding box	Extent				Canny.Mean
		Aspect ratio of a fitted ellipse	Min.Maj.Axis	**Haralick texture features**	2	Angular 2nd moment (ASM): A measure of homogeneity	Mean
		A measure of convexity	Solidity				Range
	6	Fourier simplifications of object shape.	FSD1		2	Contrast: Intensity variation for neighbouring pixels	Mean
			FSD2				Range
			FSD3		2	Correlation: Intensity correlation for neighboring pixels	Mean
			FSD4				Range
			FSD5		2	Sum of squares: A measure of variance	Mean
			FSD6				Range
Intensity	12	Nucleus hematoxylin intensity features.	Min		2	Inverse difference moment: A measure of homogeneity	Mean
			Max				Range
			Mean		4	Sum average & Sum variance for all features	Mean
			Median				Range
			MeanMed.Diff		2	Sum entropy features	Mean
			Std				Range
			IQR		2	Entropy	Mean
			MAD				Range
			Skewness		4	Difference variance & Difference entropy	Mean
			Kurtosis				Range
			HistEnergy		4	Information Measure of Correlation (IMC) (2 types)	Mean
			HistEntropy				Range

Supplementary Figures

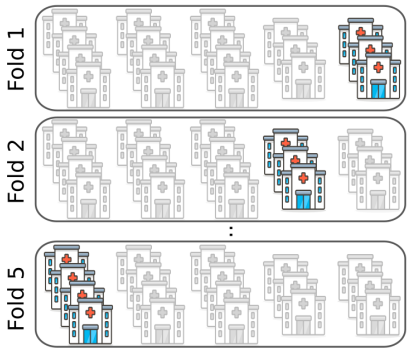

Fig. S1. Internal-external cross-validation procedure. The TCGA dataset originates from multiple institutions, and we used this fact to obtain an estimate of the external analytic validity of our models. Fold 1 was used for tuning hyper parameters, while folds 4-5 were used as external testing sets.

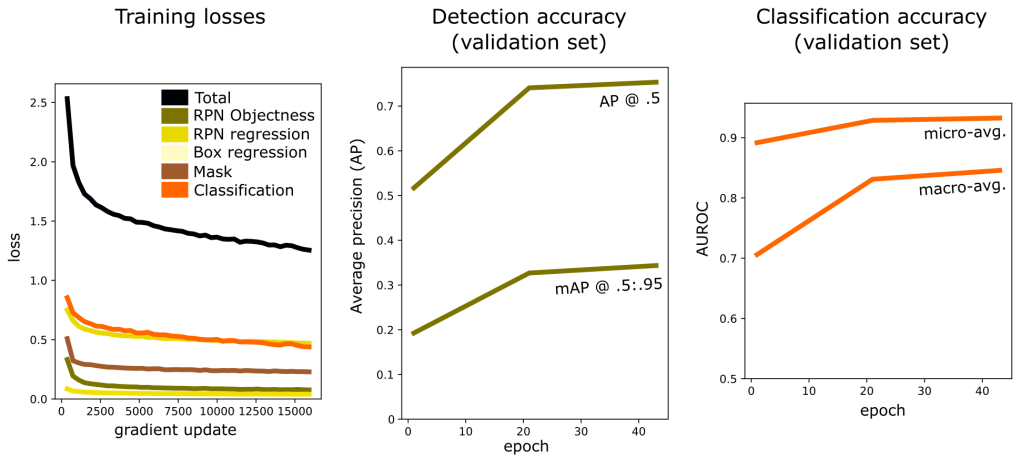

Fig. S2. Progression of NuCLS model training and convergence on fold 1. Our prototyping experiments on fold 1 (not shown) showed that the detection model started overfitting after 15k detection updates, so we froze detection weights after 15k iterations and allowed 1k extra iterations for fine-tuning of the classification layers. Abbreviations: RPN, region proposal network; AP@.5, average precision when a threshold of 0.5 is used for considering a detection to be true, mAP@.5:.95, mean average precision at a range of detection thresholds between 0.5 and 0.95; AUROC, area under receiver-operator characteristics curve.

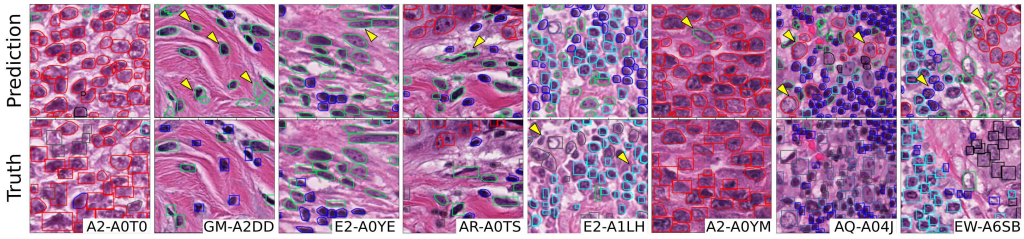

Fig. S3. Additional examples showing qualitative performance of NuCLS model on testing sets. The displayed ground truth comes from the pathologist-corrected single-rater dataset. The images are representative of a number of different hospitals in each of the testing sets from the cross-validation scheme. Detection and classification performance closely matches the ground truth, and discrepancies are marked by arrows. Not all discrepancies are algorithmic errors, including: i. adjacent nuclei that could conceivably be viewed as a single nucleus; ii. missing annotations; iii. morphologically ambiguous nuclei. Some errors arise from the lack of incorporation of contextual information in our models. Without low power context, macrophages and normal ductal/acinar cells may look morphologically similar to tumor cells.

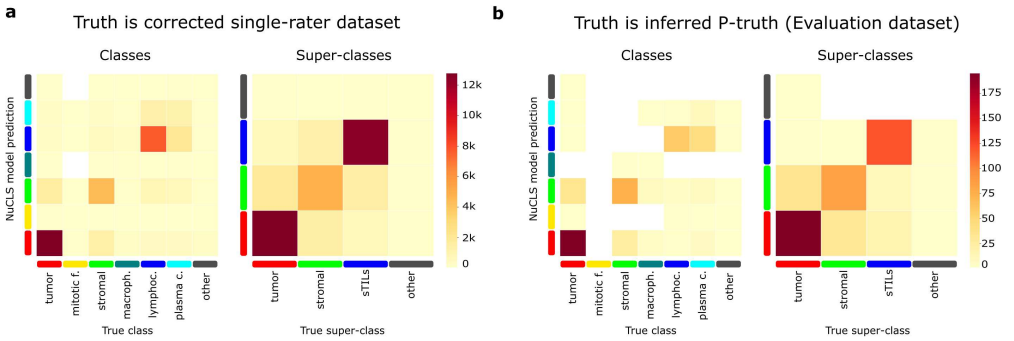

Fig. S4. Confusion matrix of NuCLS model predictions on the testing sets. For each of folds 2-5, the NuCLS model trained on the single-rater dataset training slides was used to predict FOVs from the corresponding testing set slides. The counts shown are aggregated over all testing sets. a. The single-rater dataset is considered to be the truth. b. Inferred truth from pathologists (inferred P-truth) on the multi-rater Evaluation dataset is considered to be the truth.

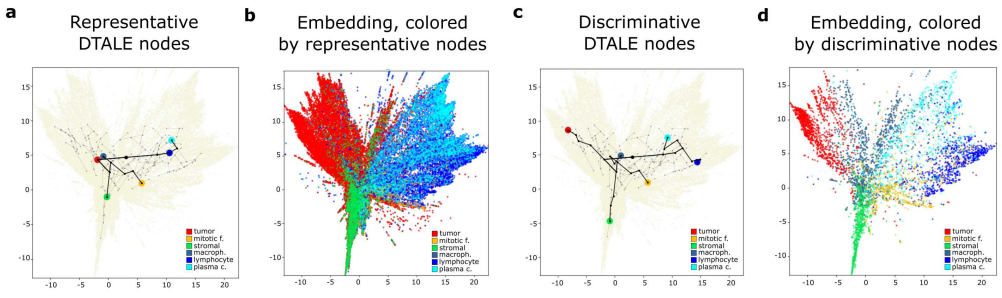

Fig. S5. Representative vs discriminative approximation of NuCLS model decisions using DTALE. a. Overlay of the full DTALE tree (light gray) on top of the embedding to which it was fitted. In black, we show paths to the nodes that allow representative approximation of NuCLS decisions, i.e. highest F-1 score. b. Nuclei that correspond to representative DTALE nodes. c. DTALE nodes that correspond to the most discriminative approximation of the NuCLS decisions, i.e. highest precision. d. Nuclei that correspond to discriminative DTALE nodes.