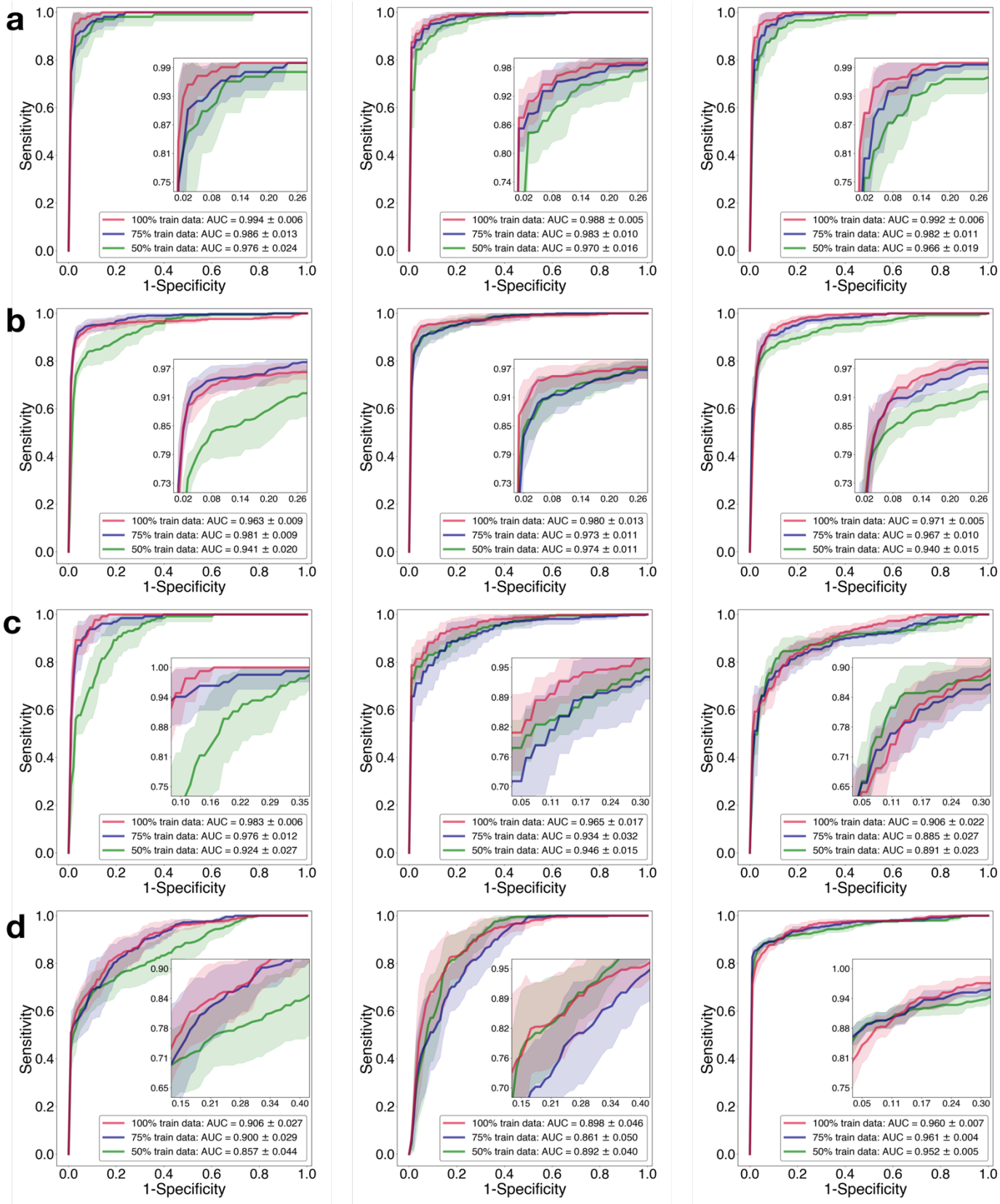
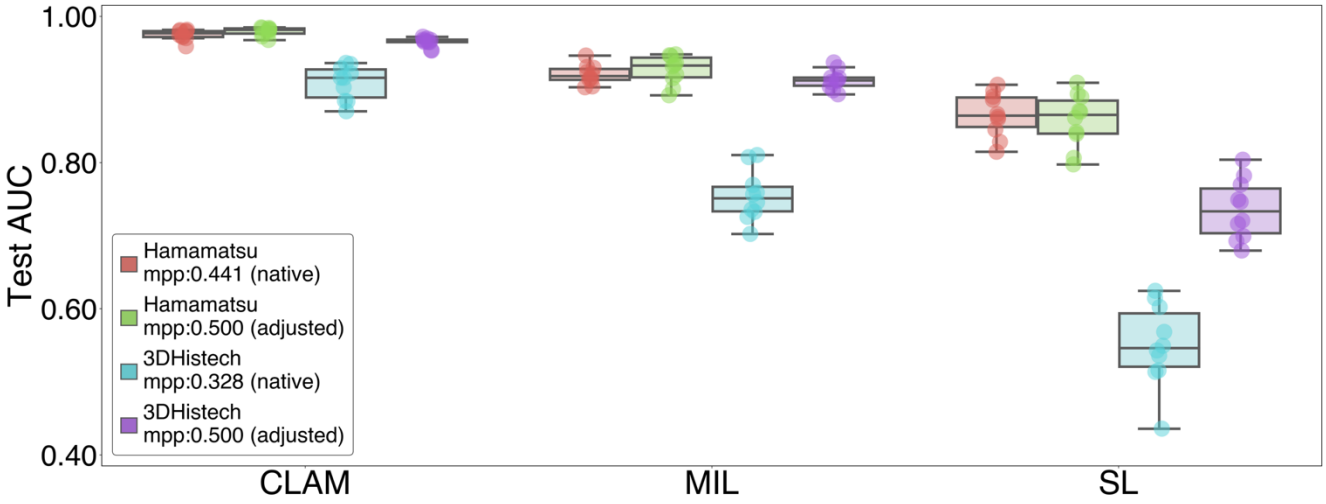
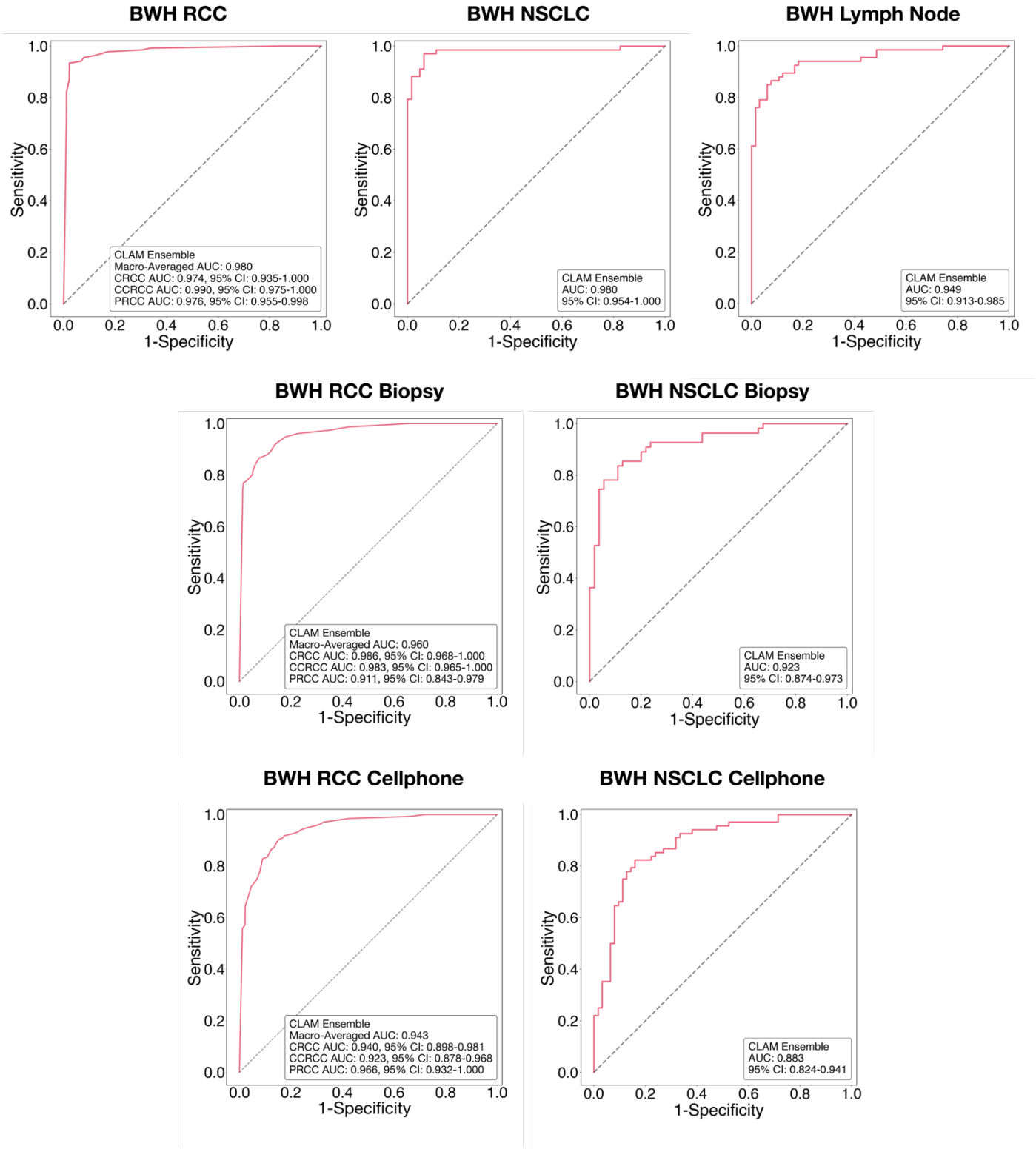
Table of contents



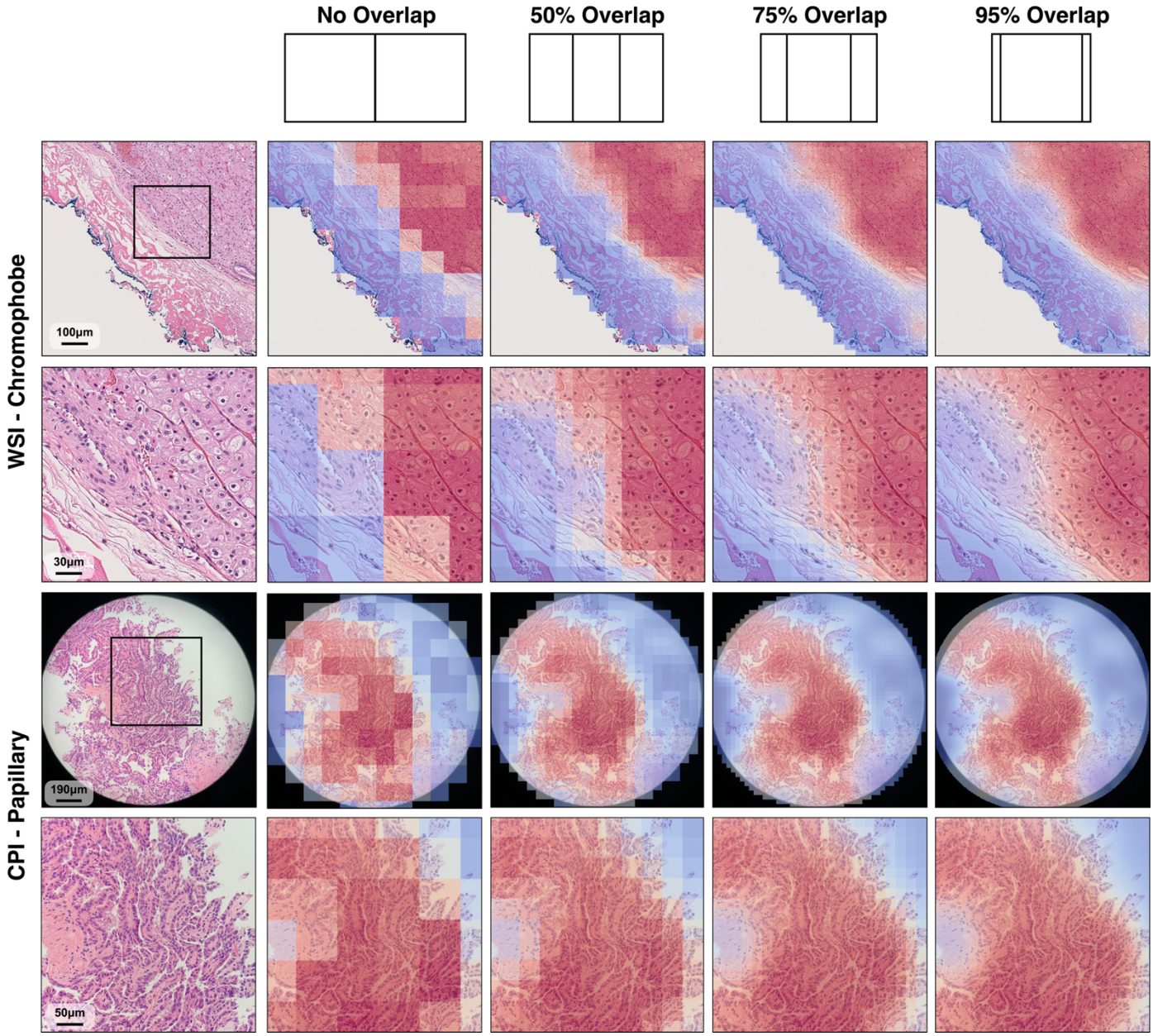
Supplementary Figure 1. Per subtype performance on RCC subtyping. For different training set sizes, all 10 CLAM models trained on the public TCGA kidney dataset are evaluated. By considering the probability predictions and ground truth labels for the 3-class classification problem as one-vs-rest (OVR), the averaged ROC curve of 10 models (confidence band shows ±1 std) is drawn for each of the three classes (**left:** chromophobe, **middle:** clear cell and **right:** papillary) on each RCC subtyping dataset: **a)** public TCGA kidney test set (n = 86), **b)** BWH independent test set (n = 135), **c)** BWH biopsy dataset (n = 92), and **d**) BWH cellphone dataset (n = 135). Area under the curve (AUC) values are shown in figure legends (± std).



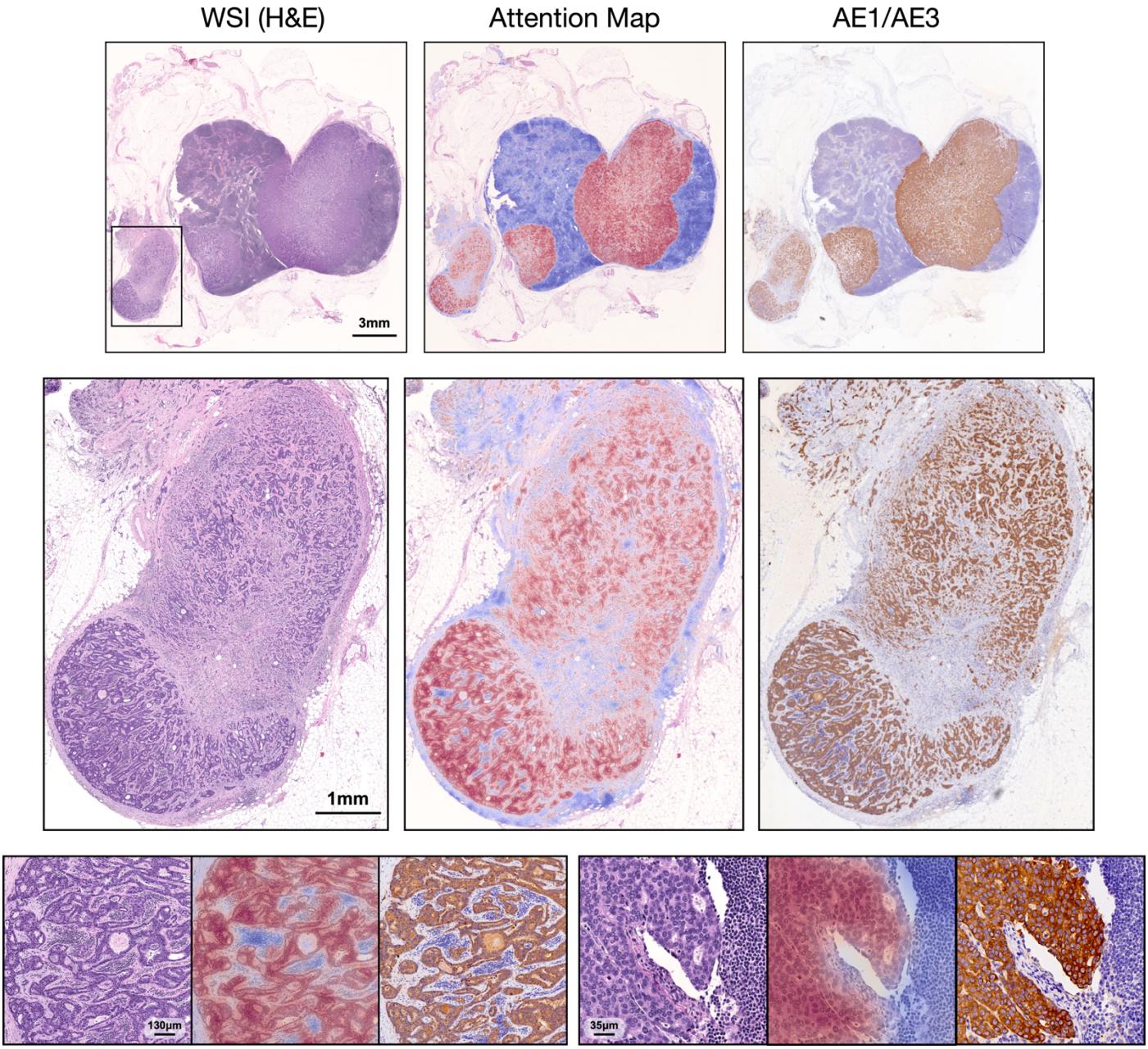
Supplementary Figure 2. Model performance on in-house NSCLC resection WSIs for different scanner hardware. Different scanner hardware produces different micron per pixel (mpp) resolution even at the same magnification. We evaluated the models trained on the public TCGA + CPTAC lung dataset (Aperio scans with an average 20x equivalent mpp of 0.50) on both WSIs (n = 131) digitized with an in-house Hamamatsu scanner (20x mpp: 0.44) and in-house 3DHistech scanner (20x mpp: 0.33). The 3DHistech scans were drastically different in terms of mpp from the training data and resulted in an average decrease of 6.5% in test AUC to 0.910 ± 0.022 from 0.975 ± 0.007. Notably, the drop in AUC performance is much larger for MIL and SL when evaluated on the 3DHistech scans, at 16.6% and 31.6% respectively. Additionally, we adjusted the mpp of our in-house scans to 0.5 by downscaling the image patches before they are embedded by the CNN feature encoder. When this simple technique is applied, the average test AUC of CLAM on the 3DHistech scans improved to 0.965 ± 0.006. These results demonstrate that CLAM is reasonably robust to technical variability introduced by different scanner hardware. The test AUC performance of all 10 trained models for each algorithm is shown for each configuration using box plot. Boxes indicate quartile values (1st, median, and 3rd) and whiskers extend to data points within 1.5x the interquartile range.



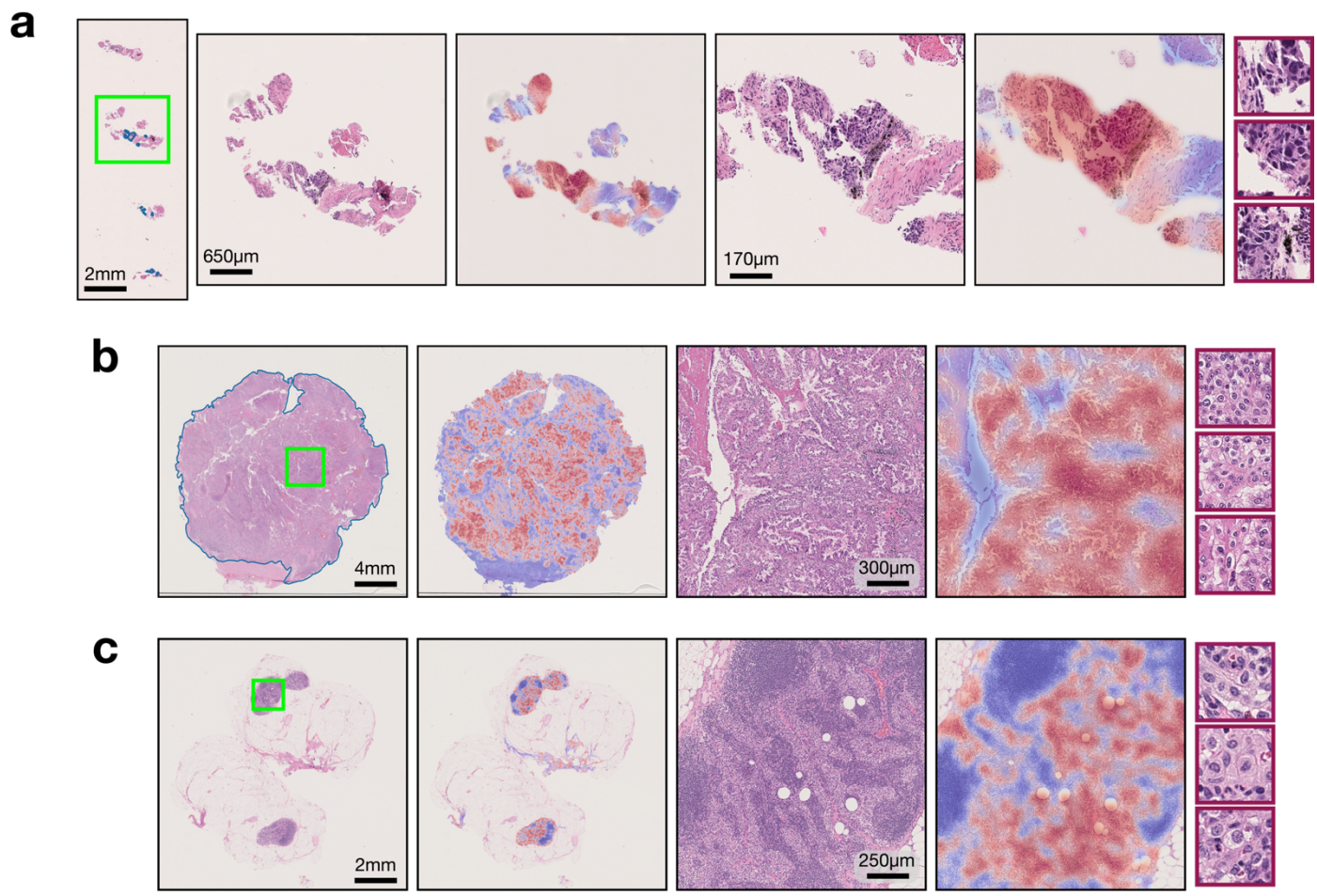
Supplementary Figure 3. Performance of CLAM ensemble system on independent test sets. For each task, we took the 10 models trained on the public datasets using 10-fold monte-carlo cross-validation (recall 80% of cases in each dataset were used to train each model) and computed their ensemble predictions by averaging the normalized probability scores over all 10 readers for each slide in the independent test set. For BWH NSCLC subtyping and axillary lymph node metastasis detection, the test AUC of the ROC curve corresponding the ensemble predictions is computed along with its 95% confidence interval (CI). For BWH RCC subtyping, the one-vs-rest AUC and its 95% CI for each subtype is computed in addition to the macro-averaged AUC. See **Supplementary Table 12 – 14** for more details.



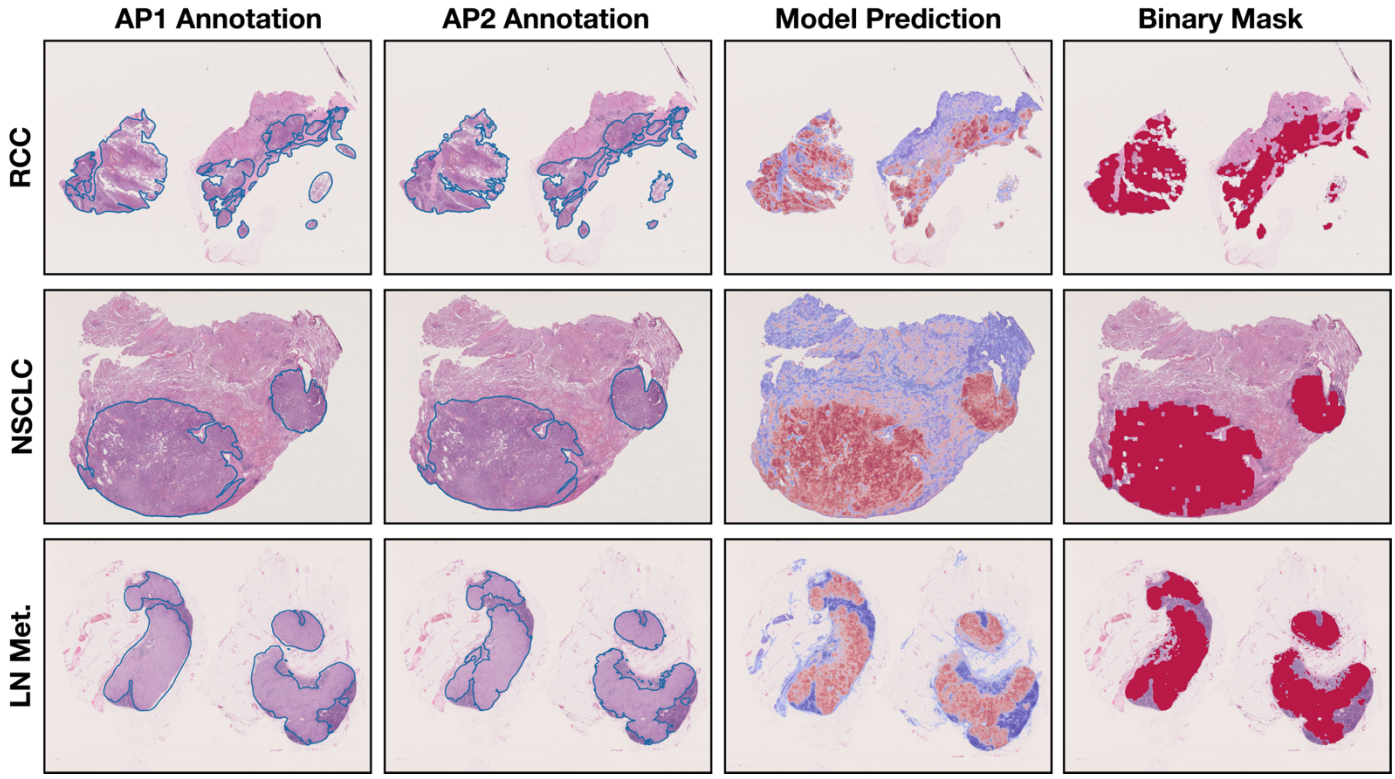
Supplementary Figure 4. Attention heatmap visualization using varying degrees of overlap. In our study, CLAM uses 256 × 256 patches to make predictions. By default, patches cover the entire extent of the tissue regions in each slide with a step size of 256 (no overlap) for fast training and inference. The resulting attention heatmap appears blocky as there can be large transitions in the attention scores assigned to neighboring patches. Instead of using interpolation techniques to estimate the attention scores for overlapped locations that are not sampled during patching, we increased the overlap between patches (up to 95% overlap) for fine-grained heatmap visualization. Attentions scores are first normalized to percentile scores by referring to the raw scores computed for all non-overlapped patch locations (this ensures that the same locations from the overlapped and non-overlapped heatmaps always have roughly the same normalized scores). Normalized attention scores are mapped to their corresponding spatial locations in the WSI and visualized (scores for overlapped regions are accumulated and averaged). As demonstrated in both the whole slide image (WSI) and cellphone image (CPI) example, using an overlap above 50% significantly reduces the blockiness of the resulting heatmap and using a 95% overlap renders the heatmap nearly completely smooth to a human observer.



Supplementary Figure 5. Validation of attention heatmap for axillary lymph node metastasis using cytokeratin (AE1/AE3) immunohistochemical staining. Subsequent slices of paraffin-embedded tissue of several positive cases of axillary lymph node metastasis are collected, cut and stained with H&E and AE1/AE3 IHC, and digitized at BWH. In the representative example, a CLAM model trained on our public lymph node metastasis training set is tested on the entire tissue region (excluding fat) of the H&E WSI using overlapping patches and a fine-grained attention heatmap corresponding to the model's prediction is created. We find that in addition to correctly detecting metastasis at the slide-level, CLAM accurately attends to metastatic regions (red in attention heatmap, gold in corresponding IHC) and often even individual tumor cells in the side-by-side comparison of the fine-grained attention heatmap and IHC-stained WSI. This promising finding suggests that while further validation is needed, in some circumstances, it might be possible to apply CLAM (which requires no pixel-level or ROI-level annotation and no special stains for training) to whole-slide-level segmentation tasks (including but not limited to predicting the corresponding IHC) that would otherwise incur either costly labor and human expertise or expensive reagents and core facilities.



Supplementary Figure 6. Analysis of misclassified cases in the independent test sets. The attention heatmaps and high attention patches can be utilized to analyze failure cases of the CLAM model. **a)** Example of squamous cell NSCLC misclassified as adenocarcinoma. Tumor regions were identified by the model and represented large, pleomorphic, poorly differentiated cells that lacked definite morphologic features or architecture of either adenocarcinoma (glandular formation, intracellular mucin, *etc.*) or squamous cell carcinoma (keratin formation, intracellular bridges, *etc.*). **b)** Example of papillary RCC misclassified as chromophobe. The model identified large, atypical, polygonal cells with either a clear or granular, eosinophilic cytoplasm in regions that did not have definitive fibrovascular cores, likely as a result of sectioning. **c)** Example of false positive misclassification in lymph node metastasis detection. The model identified larger cells with irregular nuclear contours and foamy cytoplasm, likely representing histiocytes that are commonly found within lymph nodes.



A. RCC

	Dice	IoU	κ
AP 1	0.789	0.690	0.726
AP 2	0.775	0.673	0.711

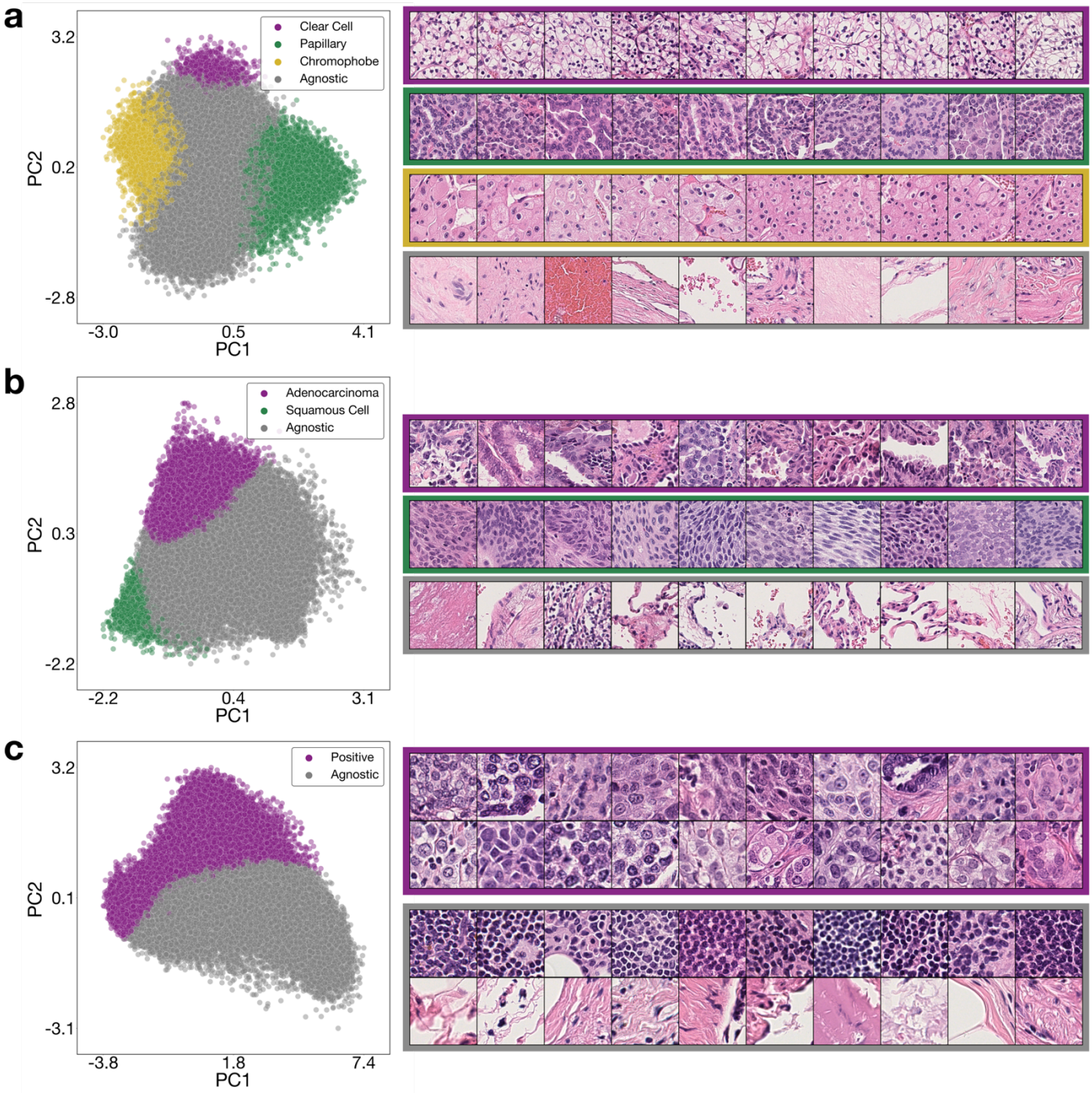
B. NSCLC

	Dice	IoU	κ
AP 1	0.777	0.651	0.733
AP 2	0.787	0.664	0.754

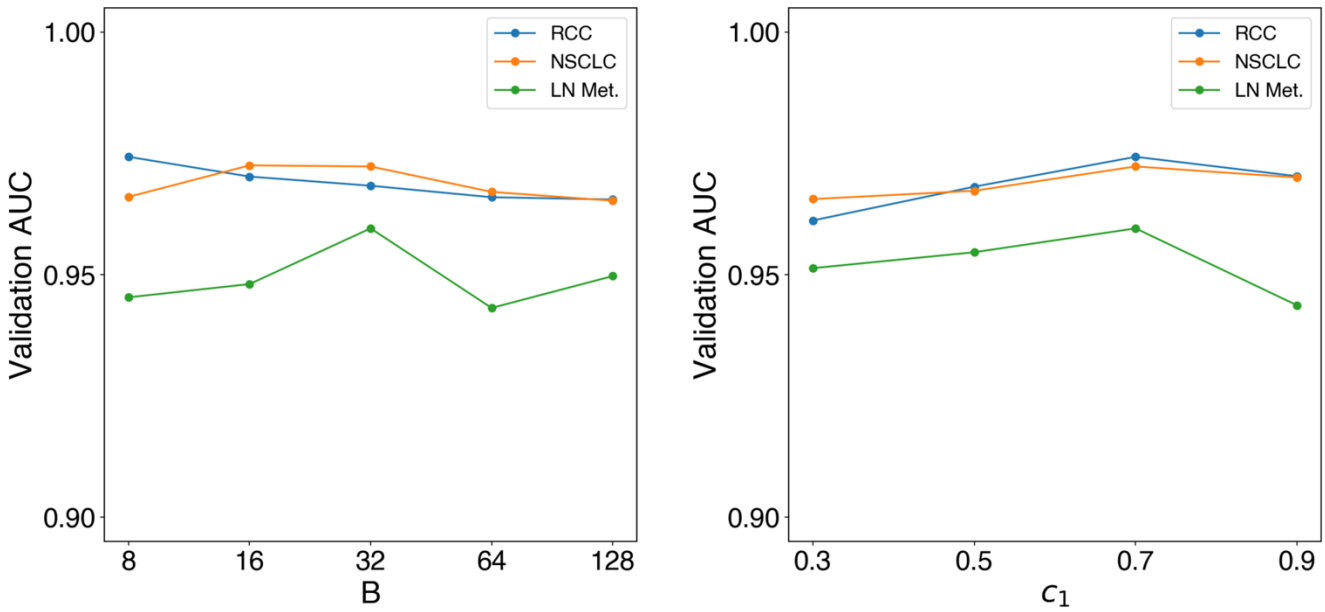
C. Lymph node met.

	Dice	IoU	κ
AP 1	0.762	0.660	0.754
AP 2	0.714	0.594	0.704

Supplementary Figure 7. Quantitative assessment of attention heatmaps using pathologist annotations. While attention heatmaps of CLAM models trained with only WSI-level labels were not intended for detailed annotation at the pixel-level, we attempted to quantitatively correlate the attention heatmaps of CLAM with tumor regions in WSIs. Two anatomic pathologists (AP) independently annotated all resection slides in our in-house RCC, NSCLC and Lymph node met. datasets. Attention heatmaps were first created by using the best performing CLAM model in terms of test AUC for each task with patches tiled at a 75% overlap and then thresholded to produce binary masks. Simple post processing techniques such as morphological closing followed by opening were used to reduce the fragmentation in the initial masks, close small holes and suppress small artifacts. The Dice score, intersection over union (IoU) and Cohen's κ were calculated against the ground truth annotations.



Supplementary Figure 8. Visualizing the patch-level feature space. To visualize the patch-level feature space, for each task, we randomly sampled 2% of patches from each slide in the independent test cohort for a total of n = 54,995, n = 64,687 and n = 136,726 patches for **a)** RCC, **b)** NSCLC and **c)** LN Met. Slides in the BWH independent test sets. We then reduced their 512-dimensional feature representations to two dimensions using PCA (**left**). For subtyping tasks (**a, b**), each patch is shaded with the class with the highest predicted probability (with $p \geq 0.5$) by the clustering layers of the model. If a patch is predicted as negative ($p < 0.5$) for all classes, it is labeled as "agnostic". We observe that patches predicted as different subtypes are separated into distinct clusters in the feature space, and patches sampled from each cluster generally exhibit morphology characteristic of each subtype. Similarly, for metastasis detection in axillary lymph nodes (**c**), patches are shaded as positive ($p \geq 0.5$ for the positive class) and agnostic ($p < 0.5$). The positive cluster generally picks out tumor cells and the agnostic cluster corresponds with immune cells and normal tissue.



Supplementary Figure 9. CLAM model performance for different hyperparameter choices. Unless otherwise specified, we used a random validation fold from our 10-fold train/val/test partitions created for each task to tune for B, which controls the number of patches to consider for the task of instance-level clustering, and then c_1 and c_2, which specify the relative contribution of the bag-level classification loss and the instance-level clustering loss in the total loss incurred for each slide during training (without loss of generality, we let $c_2 = 1 - c_1$ and tuned for different values of c_1). The models were trained using 50% of cases in the training set corresponding the selected validation set. In each task, we did not notice a large difference in the validation performance for different hyperparameter choices.

10-fold CV 100% Train Data	AUC	bACC	F1	mAP
mMIL	0.9730 ± 0.0135	0.8946 ± 0.0197	0.8932 ± 0.0262	0.9482 ± 0.0266
SL	0.9859 ± 0.0074	0.9122 ± 0.0326	0.9084 ± 0.0342	0.9740 ± 0.0130
CLAM	0.9915 ± 0.0038	0.9205 ± 0.0175	0.9212 ± 0.0264	0.9820 ± 0.0109
10-fold CV 75% Train Data				
mMIL	0.9699 ± 0.0147	0.8778 ± 0.0298	0.8742 ± 0.0331	0.9390 ± 0.0322
SL	0.9693 ± 0.0089	0.8690 ± 0.0271	0.8260 ± 0.0346	0.9491 ± 0.0177
CLAM	0.9838 ± 0.0097	0.9021 ± 0.0295	0.8853 ± 0.0260	0.9613 ± 0.0278
10-fold CV 50% Train Data				
mMIL	0.9304 ± 0.0274	0.8058 ± 0.0458	0.8030 ± 0.0488	0.8825 ± 0.0414
SL	0.9171 ± 0.0226	0.7955 ± 0.0496	0.7288 ± 0.0638	0.8788 ± 0.0290
CLAM	0.9705 ± 0.0163	0.8788 ± 0.0449	0.8679 ± 0.0470	0.9453 ± 0.0343
10-fold CV 25% Train Data				
mMIL	0.8562 ± 0.0486	0.7177 ± 0.0785	0.7063 ± 0.0807	0.7670 ± 0.0833
SL	0.8784 ± 0.0341	0.7094 ± 0.0566	0.6289 ± 0.0529	0.7838 ± 0.0575
CLAM	0.9532 ± 0.0169	0.8183 ± 0.0524	0.8156 ± 0.0507	0.9038 ± 0.0314
10-fold CV 10% Train Data				
mMIL	0.7772 ± 0.0637	0.5517 ± 0.1048	0.5464 ± 0.0985	0.6426 ± 0.1016
SL	0.8667 ± 0.0237	0.6684 ± 0.0364	0.5382 ± 0.0522	0.7608 ± 0.0383
CLAM	0.9044 ± 0.0366	0.7619 ± 0.0820	0.7497 ± 0.0625	0.8248 ± 0.0608

Supplementary Table 1. RCC subtyping: cross-validation performance on TCGA dataset. The 10-fold average performance (± std) in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) are reported (n = 86). Macro-averaging is used for one-vs-rest AUC, F1 and mAP.

10-fold CV 100% Train Data	AUC	bACC	F1	mAP
MIL	0.9062 ± 0.0258	0.8322 ± 0.0282	0.8011 ± 0.0321	0.8919 ± 0.0310
SL	0.8735 ± 0.0345	0.7738 ± 0.0406	0.7304 ± 0.0479	0.8317 ± 0.0451
CLAM	0.9561 ± 0.0179	0.8891 ± 0.0252	0.8675 ± 0.0284	0.9393 ± 0.0253
10-fold CV 75% Train Data				
MIL	0.9034 ± 0.0234	0.8332 ± 0.0330	0.8025 ± 0.0380	0.8899 ± 0.0274
SL	0.8496 ± 0.0320	0.7534 ± 0.0301	0.7094 ± 0.0343	0.8024 ± 0.0529
CLAM	0.9510 ± 0.0190	0.8919 ± 0.0288	0.8713 ± 0.0342	0.9262 ± 0.0323
10-fold CV 50% Train Data				
MIL	0.8878 ± 0.0251	0.8140 ± 0.0284	0.7800 ± 0.0324	0.8663 ± 0.0293
SL	0.7894 ± 0.0462	0.6978 ± 0.0431	0.6571 ± 0.0507	0.7290 ± 0.0696
CLAM	0.9406 ± 0.0195	0.8733 ± 0.0244	0.8510 ± 0.0275	0.9050 ± 0.0396
10-fold CV 25% Train Data				
MIL	0.7919 ± 0.0677	0.7065 ± 0.0891	0.6181 ± 0.1853	0.7415 ± 0.1054
SL	0.6901 ± 0.0508	0.6076 ± 0.0432	0.6112 ± 0.0313	0.6157 ± 0.0467
CLAM	0.9032 ± 0.0248	0.8105 ± 0.0505	0.7719 ± 0.0707	0.8640 ± 0.0380
10-fold CV 10% Train Data				
MIL	0.7134 ± 0.0728	0.6380 ± 0.0695	0.4894 ± 0.1933	0.6620 ± 0.0865
SL	0.6152 ± 0.0727	0.5643 ± 0.0543	0.5628 ± 0.0583	0.5649 ± 0.0780
CLAM	0.7983 ± 0.0632	0.7213 ± 0.0610	0.6845 ± 0.0722	0.7274 ± 0.0897

Supplementary Table 2. NSCLC subtyping: cross-validation performance on TCGA + CPTAC dataset. The 10-fold average performance (± std) in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) are reported (n = 196).

10-fold CV 100% Train Data	AUC	bACC	F1	mAP
MIL	0.9033 ± 0.1591	0.8547 ± 0.1303	0.8004 ± 0.2214	0.8891 ± 0.1680
SL	0.7041 ± 0.1053	0.5887 ± 0.0418	0.3024 ± 0.1269	0.6306 ± 0.0918
CLAM	0.9532 ± 0.0292	0.9042 ± 0.0365	0.8878 ± 0.0471	0.9464 ± 0.0302
10-fold CV 75% Train Data				
MIL	0.9047 ± 0.1132	0.8392 ± 0.1092	0.7738 ± 0.2297	0.8848 ± 0.1594
SL	0.6008 ± 0.0796	0.5296 ± 0.0258	0.1169 ± 0.0985	0.5178 ± 0.0824
CLAM	0.9298 ± 0.0437	0.8847 ± 0.0251	0.8610 ± 0.0347	0.9293 ± 0.0403
10-fold CV 50% Train Data				
MIL	0.7396 ± 0.2189	0.6851 ± 0.1701	0.4852 ± 0.3594	0.6754 ± 0.2722
SL	0.5759 ± 0.0782	0.5272 ± 0.0826	0.4829 ± 0.0994	0.4640 ± 0.1018
CLAM	0.9286 ± 0.0278	0.8717 ± 0.0290	0.8421 ± 0.0417	0.9079 ± 0.0410
10-fold CV 25% Train Data				
MIL	0.5861 ± 0.1468	0.5514 ± 0.1139	0.2887 ± 0.2299	0.5073 ± 0.1824
SL	0.5489 ± 0.0972	0.5227 ± 0.0827	0.4746 ± 0.0979	0.4523 ± 0.1252
CLAM	0.8538 ± 0.0542	0.8106 ± 0.0510	0.7535 ± 0.0700	0.8236 ± 0.0609
10-fold CV 10% Train Data				
MIL	0.5973 ± 0.1547	0.5775 ± 0.1151	0.2168 ± 0.2956	0.5014 ± 0.2030
SL	0.5384 ± 0.0964	0.5124 ± 0.0829	0.4511 ± 0.1118	0.4313 ± 0.0988
CLAM	0.8419 ± 0.0481	0.7861 ± 0.0635	0.7234 ± 0.0863	0.7978 ± 0.0828

Supplementary Table 3. Lymph node metastasis detection: cross-validation performance on Camelyon16 + Camelyon17 dataset. The 10-fold average performance (± std) in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) are reported (n = 89).

A. LN metastasis detection

10% of Train Data	AUC	bACC	F1	mAP
CLAM (no clustering)	0.7370 ± 0.1484	0.7156 ± 0.1283	0.6064 ± 0.2194	0.6914 ± 0.2020
CLAM (w/ clustering)	0.8419 ± 0.0481	0.7861 ± 0.0635	0.7234 ± 0.0863	0.7978 ± 0.0828
50% of Train Data				
CLAM (no clustering)	0.9182 ± 0.0398	0.8565 ± 0.0333	0.8234 ± 0.0518	0.8961 ± 0.0499
CLAM (w/ clustering)	0.9286 ± 0.0278	0.8717 ± 0.0290	0.8421 ± 0.0417	0.9079 ± 0.0410
100% of Train Data				
CLAM (no clustering)	0.9405 ± 0.0425	0.8946 ± 0.0292	0.8763 ± 0.0345	0.9404 ± 0.0320
CLAM (w/ clustering)	0.9532 ± 0.0292	0.9042 ± 0.0365	0.8878 ± 0.0471	0.9464 ± 0.0302

B. NSCLC subtyping

10% of Train Data	AUC	bACC	F1	mAP
CLAM (no clustering)	0.7788 ± 0.0674	0.6850 ± 0.0927	0.5592 ± 0.2399	0.6992 ± 0.0806
CLAM (w/ clustering)	0.7983 ± 0.0632	0.7213 ± 0.0610	0.6845 ± 0.0722	0.7274 ± 0.0897
50% of Train Data				
CLAM (no clustering)	0.9300 ± 0.0259	0.8576 ± 0.0242	0.8318 ± 0.0300	0.8948 ± 0.0525
CLAM (w/ clustering)	0.9406 ± 0.0195	0.8733 ± 0.0244	0.8510 ± 0.0275	0.9050 ± 0.0396
100% of Train Data				
CLAM (no clustering)	0.9523 ± 0.0158	0.8886 ± 0.0211	0.8662 ± 0.0235	0.9270 ± 0.0238
CLAM (w/ clustering)	0.9561 ± 0.0179	0.8891 ± 0.0252	0.8675 ± 0.0284	0.9393 ± 0.0253

C. RCC subtyping

10% of Train Data	AUC	bACC	F1	mAP
CLAM (no clustering)	0.8902 ± 0.0337	0.7284 ± 0.0970	0.7152 ± 0.0995	0.8120 ± 0.0509
CLAM (w/ clustering)	0.9044 ± 0.0366	0.7619 ± 0.0820	0.7497 ± 0.0625	0.8248 ± 0.0608
50% of Train Data				
CLAM (no clustering)	0.9660 ± 0.0200	0.8633 ± 0.0633	0.8513 ± 0.0599	0.9360 ± 0.0363
CLAM (w/ clustering)	0.9705 ± 0.0163	0.8788 ± 0.0449	0.8679 ± 0.0470	0.9453 ± 0.0343
100% of Train Data				
CLAM (no clustering)	0.9890 ± 0.0078	0.9219 ± 0.0261	0.9212 ± 0.0301	0.9766 ± 0.0190
CLAM (w/ clustering)	0.9915 ± 0.0038	0.9205 ± 0.0175	0.9212 ± 0.0264	0.9820 ± 0.0109

Supplementary Table 4. Ablation experiments. Experiments for CLAM with and without the clustering constraint were performed on the public dataset partitions (the same as our main study described by **Figure 2**) for all 3 disease models and different sized subsets of the full training set. The 10-fold mean test performance (± std) is reported for the AUC, bACC (balanced accuracy), F1 and mAP (mean average precision) score. For multi-class RCC subtyping, macro-averaging was used for all metrics to account for class imbalance. Clustering improves performance for smaller data denominations, particularly for more difficult tasks such as LN metastasis detection. Clustering also offers a mechanism of enhanced interpretability shown in **Supplementary Figure 8.**

A. LN metastasis detection

	AUC	bACC	F1	mAP
Train/Val/Test: 40/10/50 (n=452)				
MIL	0.7154 ± 0.1836	0.6509 ± 0.1681	0.4183 ± 0.3416	0.6069 ± 0.2609
SL	0.6589 ± 0.0306	0.5967 ± 0.0285	0.4676 ± 0.0369	0.5244 ± 0.0383
CLAM	0.9199 ± 0.0135	0.8665 ± 0.0232	0.8362 ± 0.0284	0.9111 ± 0.0176
Train/Val/Test: 60/10/30 (n=269)				
MIL	0.7628 ± 0.1682	0.6893 ± 0.1661	0.4918 ± 0.3412	0.6720 ± 0.2451
SL	0.6944 ± 0.0476	0.5916 ± 0.0351	0.3330 ± 0.1064	0.5934 ± 0.0694
CLAM	0.9374 ± 0.0238	0.8839 ± 0.0382	0.8574 ± 0.0476	0.9290 ± 0.0237

B. NSCLC subtyping

	AUC	bACC	F1	mAP
Train/Val/Test: 40/10/50 (n=984)				
MIL	0.9043 ± 0.0088	0.8217 ± 0.0147	0.7867 ± 0.0170	0.8763 ± 0.0102
SL	0.8452 ± 0.0327	0.7654 ± 0.0317	0.7226 ± 0.0370	0.7725 ± 0.0540
CLAM	0.9424 ± 0.0058	0.8635 ± 0.0178	0.8371 ± 0.0210	0.9116 ± 0.0061
Train/Val/Test: 60/10/30 (n=598)				
MIL	0.9099 ± 0.0147	0.8247 ± 0.0197	0.7908 ± 0.0256	0.8873 ± 0.0206
SL	0.8597 ± 0.0303	0.7899 ± 0.0261	0.7531 ± 0.0294	0.8054 ± 0.0424
CLAM	0.9482 ± 0.0109	0.8740 ± 0.0134	0.8508 ± 0.0161	0.9220 ± 0.0135

C. RCC subtyping

	AUC	bACC	F1	mAP
Train/Val/Test: 40/10/50 (n=442)				
MIL	0.9679 ± 0.0045	0.8807 ± 0.0162	0.8768 ± 0.0095	0.9356 ± 0.0106
SL	0.9745 ± 0.0080	0.8823 ± 0.0188	0.8794 ± 0.0178	0.9521 ± 0.0140
CLAM	0.9775 ± 0.0074	0.8831 ± 0.0207	0.8849 ± 0.0196	0.9529 ± 0.0139
Train/Val/Test: 60/10/30 (n=262)				
MIL	0.9717 ± 0.0044	0.8930 ± 0.0255	0.8877 ± 0.0190	0.9396 ± 0.0100
SL	0.9783 ± 0.0065	0.8876 ± 0.0277	0.8854 ± 0.0271	0.9530 ± 0.0221
CLAM	0.9792 ± 0.0115	0.8946 ± 0.0202	0.8893 ± 0.0218	0.9573 ± 0.0171

Supplementary Table 5. Performance comparison with weakly-supervised baseline methods on public datasets using additional dataset partitions. Additional experiments for comparing the performance of CLAM with weakly-supervised baseline algorithms trained using reduced data were performed on public datasets for all 3 disease models. Specifically, 40/10/50 and 60/10/30 train/val/test partitions were investigated on the same datasets used in our main cross-validation study (**Figure 2**)**.** These experiments have the additional benefit of allowing the algorithms to be assessed on larger held-out test sets (compared to *i.e.,* using 10% of the dataset as the test set). Consistent with the rest of our study, we used repeated 10-fold partitions for each task and the mean test performance (± std) is reported for the AUC, bACC (balanced accuracy), F1 and mAP (mean average precision) score. For multi-class RCC subtyping, macro-averaging was used for all metrics to account for class imbalance. We used the same hyperparameters for CLAM as in the main study and ablation experiments.

A. LN metastasis detection: Train C17 → Test C16 (n=399)

25% of Train Data	AUC	bACC	F1	mAP
MIL	0.6220 ± 0.1455	0.5645 ± 0.1012	0.2830 ± 0.2788	0.5402 ± 0.1834
SL	0.5726 ± 0.0206	0.5083 ± 0.0057	0.0445 ± 0.0378	0.5277 ± 0.0277
CLAM	0.8101 ± 0.1496	0.7597 ± 0.1241	0.6496 ± 0.2727	0.7860 ± 0.1696
100% of Train Data				
MIL	0.8451 ± 0.1503	0.7598 ± 0.1181	0.6836 ± 0.1905	0.8337 ± 0.1716
SL	0.6089 ± 0.0192	0.5104 ± 0.0016	0.0415 ± 0.0078	0.5661 ± 0.0208
CLAM	0.9120 ± 0.0118	0.8416 ± 0.0284	0.8100 ± 0.0408	0.9201 ± 0.0093

B. NSCLC subtyping: Train TCGA → Test CPTAC (n=974)

25% of Train Data	AUC	bACC	F1	mAP
MIL	0.7919 ± 0.0677	0.7065 ± 0.0891	0.6181 ± 0.1853	0.7415 ± 0.1054
SL	0.6901 ± 0.0508	0.6076 ± 0.0432	0.6112 ± 0.0313	0.6157 ± 0.0467
CLAM	0.9032 ± 0.0248	0.8105 ± 0.0505	0.7719 ± 0.0707	0.8640 ± 0.0380
100% of Train Data				
MIL	0.9062 ± 0.0258	0.8322 ± 0.0282	0.6836 ± 0.1905	0.8337 ± 0.1716
SL	0.8735 ± 0.0345	0.7738 ± 0.0406	0.7304 ± 0.0479	0.8317 ± 0.0451
CLAM	0.9557 ± 0.0199	0.8825 ± 0.0289	0.8604 ± 0.0338	0.9358 ± 0.0278

C. RCC subtyping: Train TCGA → Test TCGA (independent sites, n=140)

25% of Train Data	AUC	bACC	F1	mAP
mMIL	0.8562 ± 0.0486	0.7177 ± 0.0785	0.7063 ± 0.0807	0.7670 ± 0.0833
SL	0.8784 ± 0.0341	0.7094 ± 0.0566	0.6289 ± 0.0529	0.7838 ± 0.0575
CLAM	0.9532 ± 0.0169	0.8183 ± 0.0524	0.8156 ± 0.0507	0.9038 ± 0.0314
100% of Train Data				
mMIL	0.9730 ± 0.0135	0.8946 ± 0.0197	0.8932 ± 0.0262	0.9482 ± 0.0266
SL	0.9859 ± 0.0074	0.9122 ± 0.0326	0.9084 ± 0.0342	0.9740 ± 0.0130
CLAM	0.9915 ± 0.0038	0.9205 ± 0.0175	0.9212 ± 0.0264	0.9820 ± 0.0109

Supplementary Table 6. Additional performance comparison with weakly-supervised baseline methods on public datasets. For NSCLC subtyping and lymph node met. detection, model development was performed using one dataset and evaluated on a separate dataset (*e.g.* train on TCGA data and test on CPTAC data). For RCC subtyping, 3 tissue source sites were selected to form an independent test set of 140 WSIs (19 Chromophobe, 23 Papillary, 98 Clear Cell) and the remaining of the TCGA dataset was used for model development. For each disease model, the data used for model development were randomly divided into a training (90% of cases) and validation (10% cases) set. Consistent with the rest of our study, we used a 10-fold partition for each task and the mean test performance (± std) is reported for the AUC, bACC (balanced accuracy), F1 and mAP (mean average precision) score. For multi-class RCC subtyping, macro-averaging was used for all metrics to account for class imbalance. We used the same hyperparameters for CLAM as in the main study and ablation experiments.

RCC Subtyping (TCGA)			
Study	**Method Name**	**Performance**	**Notes**
Tabibu et al. 2019 [1]	Resnet-34 + DAG-SVM	AUC: 0.93 @20X on TCGA	
Ours	CLAM	AUC: 0.991 @20X on TCGA	
NSCLC Subtyping (TCGA)			
Wang et al. 2019 [2]	CNN-AvgFea-Norm3-based RF	AUC: 0.856 @20X on TCGA	
Xu et al. 2015 [3]	Pretrained-Feature-Norm3	AUC: 0.832 @20X on TCGA	
Hou et al. 2016 [4]	EM-CNN-Fea-SVM	AUC: 0.816 @20X on TCGA	
Yu et al. 2016 [5]	SVM (Gaussian Kernel)	AUC: 0.75 @40X on TCGA	
Khosravi et al. 2018 [6]	Inception-V1 CNN	AUC: 0.89 @40X on TCGA	
Coudray et al. 2018 [7]	Inception-V4 CNN SL	AUC: 0.95 @20X on TCGA	LUAD vs LUSC (without normal)
Ours	CLAM	AUC: 0.963 @20X on TCGA	
LN Met. Detection (Camelyon 16)			
Campanella et al. 2019 [8]	MIL-RNN	AUC: 0.899 @20X on C16 Test	Weakly supervised, trained on in-house data
Tellez et al. 2019 [9]	Neural image compression	AUC: 0.704 @20X on C16 Test	Weakly supervised
Ours	CLAM	AUC: 0.936 @40X on C16 Test	Weakly supervised
Chen et al. 2016 [10]	CNN-maxpool	AUC: 0.942 @40X on C16 Test	Uses pixel-level annotation (fully supervised)
Koohbanani et al. 2018 [11]	Multi-resolution CNN ensemble	AUC: 0.990 @40X + 20X on C16 Test	Uses pixel-level annotation (fully supervised)
Wang et al. 2016 [10]	CNN-RF	AUC: 0.994 @40X on C16 Test	Uses pixel-level annotation (fully supervised)
Zhong et al. 2016 [10]	CNN-RF	AUC: 0.976 @40X on C16 Test	Uses pixel-level annotation (fully supervised)

[1] Tabibu, Sairam, P. K. Vinod, and C. V. Jawahar. "Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning." Scientific reports 9.1 (2019): 1-9.
[2] Wang, Xi, et al. "Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis." IEEE Transactions on Cybernetics (2019).
[3] Xu, Yan, et al. "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation." 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015.
[4] Hou, Le, et al. "Patch-based convolutional neural network for whole slide tissue image classification." Proceedings of the ieee conference on computer vision and pattern recognition. 2016.
[5] Yu, Kun-Hsing, et al. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." Nature communications 7.1 (2016): 1-10.
[6] Khosravi, Pegah, et al. "Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images." EBioMedicine 27 (2018): 317-328.
[7] Coudray, Nicolas, et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." Nature medicine 24.10 (2018): 1559-1567.
[8] Campanella, Gabriele, et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images." Nature medicine 25.8 (2019): 1301-1309.
[9] Tellez, David, et al. "Neural image compression for gigapixel histopathology image analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).
[10] Bejnordi, Babak Ehteshami, et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." Jama 318.22 (2017): 2199-2210.
[11] Koohbanani, Navid Alemi, et al. "Significance of hyperparameter optimization for metastasis detection in breast histology images." Computational Pathology and Ophthalmic Medical Image Analysis. Springer, Cham, 2018. 139-147.

Supplementary Table 7. Performance reported by related works. For the Camelyon16 challenge, only the top 3 performing algorithms from the official leader board are included, full leader board can be accessed at: https://camelyon16.grand-challenge.org/Results/.

Dataset	Source	Total Slides	Number of Slides Per Class	Scan Magnification	Scanner	Process Magnification	Average Number of Patches Per Slide
Public Datasets (Used for Training, Validation, Test)							
TCGA Kidney: RCC Subtyping	TCGA, Public	884	CRCC: 111, CCRCC: 489, PRCC: 284	40X or 20X	Aperio (exact model and mpp varies)	20X	13907
TCGA + CPTAC Lung: NSCLC Subtyping	TCGA and CPTAC, Public	1967	LUAD: 1175, LUSC: 792	40X or 20X	Aperio (exact model and mpp varies)	20X	9958
C16 + C17: Lymph Node Metastasis Detection	Camelyon16 and 17, Public	899	Negative: 591, Positive: 308	40X	Varies	40X	41802
Independent Test Cohorts (Only for Testing)							
BWH Kidney: RCC Subtyping	BWH, Internal	135	CRCC: 43, CCRCC: 46, PRCC: 46	40X	Hamamatsu S210	20X	20394
BWH Kidney: RCC Subtyping (Biopsy)	BWH, Internal	92	CRCC: 13, CCRCC: 53, PRCC: 26	40X	Hamamatsu S210	20X	1709
BWH Kidney: RCC Subtyping (Cellphone Imaged)	BWH, Internal	135	CRCC: 43, CCRCC: 46, PRCC: 46	20X Objective, 10X Eyepiece	iPhone X on Olympus Microscope	20X	419
BWH Lung: NSCLC Subtyping (Resection)	BWH, Internal	131	LUAD: 63, LUSC: 68	40X(Hamamatsu), 20X (3DH)	Hamamatsu S210 & 3DHistech Mirax 150	20X	24714
BWH Lung: NSCLC Subtyping (Biopsy)	BWH, Internal	110	LUAD: 55, LUSC: 55	40X	Hamamatsu S210	20X	820
BWH Lung: NSCLC Subtyping (Cellphone Imaged)	BWH, Internal	131	LUAD: 63, LUSC: 68	20X Objective (Olympus)	iPhone X on Olympus Microscope	20X	406
BWH Axillary Lymph Node: Lymph Node Metastasis Detection	BWH, Internal	133	Negative: 66, Positive: 67	40X	Hamamatsu S210	20X	51426

Supplementary Table 8. Dataset summary. Summary of all datasets used in the study.

10-fold CV 100% Train Data	AUC	bACC	F1	mAP
mMIL	0.9616 ± 0.0123	0.8834 ± 0.0203	0.8841 ± 0.0197	0.9415 ± 0.0158
SL	0.9371 ± 0.0154	0.4982 ± 0.0467	0.4263 ± 0.0605	0.9003 ± 0.0367
CLAM	0.9716 ± 0.0082	0.8916 ± 0.0247	0.8935 ± 0.0236	0.9603 ± 0.0103
10-fold CV 75% Train Data				
mMIL	0.9152 ± 0.0203	0.7928 ± 0.0515	0.7803 ± 0.0628	0.8782 ± 0.0282
SL	0.9463 ± 0.0091	0.6863 ± 0.0676	0.6498 ± 0.0891	0.9147 ± 0.0132
CLAM	0.9734 ± 0.0076	0.7941 ± 0.0477	0.7759 ± 0.0602	0.9596 ± 0.0110
10-fold CV 50% Train Data				
mMIL	0.8987 ± 0.0332	0.7894 ± 0.0517	0.7849 ± 0.0557	0.8678 ± 0.0400
SL	0.9136 ± 0.0175	0.6143 ± 0.1320	0.5643 ± 0.1745	0.8716 ± 0.0234
CLAM	0.9518 ± 0.0113	0.7834 ± 0.0532	0.7783 ± 0.0585	0.9282 ± 0.0164
10-fold CV 25% Train Data				
mMIL	0.8121 ± 0.0918	0.7013 ± 0.0971	0.6955 ± 0.1063	0.7573 ± 0.1081
SL	0.8813 ± 0.0095	0.5075 ± 0.0275	0.4177 ± 0.0361	0.8145 ± 0.0167
CLAM	0.9545 ± 0.0168	0.8316 ± 0.0439	0.8315 ± 0.0462	0.9263 ± 0.0267
10-fold CV 10% Train Data				
mMIL	0.7911 ± 0.0643	0.5551 ± 0.0740	0.4992 ± 0.0895	0.7037 ± 0.0788
SL	0.8466 ± 0.0175	0.5290 ± 0.0555	0.4457 ± 0.0800	0.7588 ± 0.0285
CLAM	0.9260 ± 0.0186	0.7847 ± 0.0222	0.7836 ± 0.0241	0.8756 ± 0.0287

Supplementary Table 9. RCC subtyping performance evaluated on the BWH RCC independent test set. The 10-fold average performance (± std) in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) are reported (n = 135). Macro-averaging is used for one-vs-rest AUC, F1 and mAP.

10-fold CV 100% Train Data	AUC	bACC	F1	mAP
MIL	0.9201 ± 0.0125	0.7426 ± 0.0543	0.6529 ± 0.1124	0.9326 ± 0.0094
SL	0.8656 ± 0.0286	0.5674 ± 0.0662	0.2220 ± 0.1913	0.8799 ± 0.0238
CLAM	0.9749 ± 0.0067	0.7551 ± 0.0719	0.6640 ± 0.1292	0.9820 ± 0.0053
10-fold CV 75% Train Data				
MIL	0.9193 ± 0.0092	0.7099 ± 0.0496	0.5894 ± 0.1081	0.9307 ± 0.0115
SL	0.9027 ± 0.0148	0.5125 ± 0.0211	0.0457 ± 0.0753	0.9158 ± 0.0135
CLAM	0.9697 ± 0.0128	0.7579 ± 0.0584	0.6744 ± 0.1028	0.9770 ± 0.0086
10-fold CV 50% Train Data				
MIL	0.9256 ± 0.0051	0.6998 ± 0.0444	0.5683 ± 0.0966	0.9381 ± 0.0044
SL	0.8498 ± 0.0440	0.5074 ± 0.0033	0.0289 ± 0.0128	0.8615 ± 0.0417
CLAM	0.9685 ± 0.0087	0.7944 ± 0.0406	0.7414 ± 0.0633	0.9722 ± 0.0110
10-fold CV 25% Train Data				
MIL	0.8577 ± 0.0515	0.6866 ± 0.0707	0.5375 ± 0.1887	0.8793 ± 0.0534
SL	0.7496 ± 0.0821	0.5527 ± 0.0581	0.2210 ± 0.2237	0.7463 ± 0.0839
CLAM	0.9154 ± 0.0268	0.7377 ± 0.0661	0.6387 ± 0.1282	0.9310 ± 0.0226
10-fold CV 10% Train Data				
MIL	0.7819 ± 0.0734	0.5804 ± 0.0432	0.3119 ± 0.1649	0.7862 ± 0.0856
SL	0.5923 ± 0.0530	0.5096 ± 0.0062	0.1510 ± 0.1131	0.5825 ± 0.0389
CLAM	0.8570 ± 0.0690	0.6255 ± 0.0663	0.4102 ± 0.1584	0.8622 ± 0.0851

Supplementary Table 10. NSCLC subtyping performance evaluated on the BWH NSCLC independent test set. The 10-fold average performance (± std) in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) are reported (n = 131).

10-fold CV 100% Train Data	AUC	bACC	F1	mAP
MIL	0.8741 ± 0.1115	0.7872 ± 0.1000	0.7554 ± 0.2255	0.8978 ± 0.1096
SL	0.7373 ± 0.0222	0.6158 ± 0.0433	0.4267 ± 0.1583	0.7599 ± 0.0218
CLAM	0.9404 ± 0.0148	0.8707 ± 0.0187	0.8707 ± 0.0203	0.9535 ± 0.0098
10-fold CV 75% Train Data				
MIL	0.8532 ± 0.1238	0.7678 ± 0.0987	0.7273 ± 0.2443	0.8839 ± 0.1168
SL	0.7444 ± 0.0080	0.5254 ± 0.0313	0.0901 ± 0.1091	0.7747 ± 0.0138
CLAM	0.9217 ± 0.0139	0.8282 ± 0.0290	0.8372 ± 0.0267	0.9289 ± 0.0128
10-fold CV 50% Train Data				
MIL	0.6964 ± 0.1786	0.6481 ± 0.1328	0.4465 ± 0.3571	0.7290 ± 0.1735
SL	0.6130 ± 0.0389	0.5008 ± 0.0054	0.6686 ± 0.0027	0.6458 ± 0.0427
CLAM	0.8754 ± 0.0146	0.7991 ± 0.0301	0.8055 ± 0.0222	0.8899 ± 0.0169
10-fold CV 25% Train Data				
MIL	0.5434 ± 0.1712	0.5491 ± 0.1131	0.2770 ± 0.3457	0.5726 ± 0.1707
SL	0.5389 ± 0.0197	0.5070 ± 0.0245	0.5671 ± 0.1809	0.6092 ± 0.0325
CLAM	0.8360 ± 0.0133	0.7598 ± 0.0206	0.7482 ± 0.0139	0.8680 ± 0.0145
10-fold CV 10% Train Data				
MIL	0.5816 ± 0.1665	0.5583 ± 0.0877	0.2174 ± 0.2947	0.6192 ± 0.1532
SL	0.4991 ± 0.0360	0.4935 ± 0.0148	0.5228 ± 0.1713	0.5527 ± 0.0486
CLAM	0.8142 ± 0.0328	0.7425 ± 0.0393	0.7257 ± 0.0370	0.8462 ± 0.0233

Supplementary Table 11. Lymph node metastasis detection performance evaluated on the BWH lymph node independent test set. The 10-fold average performance (± std) in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) are reported (n = 133).

BWH Independent Test Set (n=135)							
Training Set Size	**AUC**	**mAP**	**F1**	**bACC**	**CRCC AUC**	**CCRCC AUC**	**PRCC AUC**
100% of Train Data	0.9800	0.9715	0.9058	0.9038	0.9740 (0.9351 - 1.0000)	0.9897 (0.9746 - 1.0000)	0.9763 (0.9550 - 0.9976)
75% of Train Data	0.9805	0.9699	0.7870	0.8038	0.9853 (0.9703 - 1.0000)	0.9844 (0.9668 - 1.0000)	0.9719 (0.9457 - 0.9981)
50% of Train Data	0.9636	0.9462	0.7910	0.7941	0.9537 (0.9219 - 0.9856)	0.9861 (0.9682 - 1.0000)	0.9509 (0.9117 - 0.9901)
25% of Train Data	0.9655	0.9459	0.8737	0.8728	0.9550 (0.9216 - 0.9884)	0.9724 (0.9460 - 0.9988)	0.9690 (0.9428 - 0.9952)
10% of Train Data	0.9441	0.9139	0.8593	0.8598	0.9671 (0.9342 - 1.0000)	0.9328 (0.8881 - 0.9776)	0.9323 (0.8806 - 0.9841)
BWH Independent Test Set (Cellphone, n=135)							
100% of Train Data	0.9427	0.9031	0.6233	0.6271	0.9398 (0.8985 - 0.9812)	0.9226 (0.8776 - 0.9676)	0.9658 (0.9316 - 1.0000)
75% of Train Data	0.9234	0.8798	0.5505	0.5676	0.9363 (0.8911 - 0.9815)	0.8710 (0.8131 - 0.9289)	0.9629 (0.9259 - 0.9999)
50% of Train Data	0.9214	0.8714	0.6333	0.6520	0.9070 (0.8515 - 0.9624)	0.9023 (0.8529 - 0.9517)	0.9548 (0.9109 - 0.9987)
25% of Train Data	0.9444	0.9088	0.7168	0.7332	0.9406 (0.8973 - 0.9839)	0.9399 (0.9025 - 0.9774)	0.9526 (0.9071 - 0.9981)
10% of Train Data	0.9274	0.8844	0.7033	0.7021	0.9477 (0.9083 - 0.9870)	0.9165 (0.8709 - 0.9620)	0.9179 (0.8541 - 0.9817)
BWH Independent Test Set (Biopsy, n=92)							
100% of Train Data	0.9599	0.9242	0.7590	0.8358	0.9864 (0.9680 - 1.0000)	0.9826 (0.9645 - 1.0000)	0.9108 (0.8431 - 0.9786)
75% of Train Data	0.9468	0.9285	0.6468	0.7153	0.9922 (0.9794 - 1.0000)	0.9497 (0.9116 - 0.9878)	0.8986 (0.8224 - 0.9748)
50% of Train Data	0.9405	0.8678	0.6873	0.7463	0.9503 (0.9081 - 0.9926)	0.9637 (0.9325 - 0.9949)	0.9073 (0.8293 - 0.9854)
25% of Train Data	0.9581	0.9121	0.7564	0.7646	0.9708 (0.9410 - 1.0000)	0.9468 (0.9062 - 0.9874)	0.9569 (0.9180 - 0.9958)
10% of Train Data	0.9429	0.8832	0.7315	0.6991	0.9669 (0.9309 - 1.0000)	0.9434 (0.8953 - 0.9915)	0.9184 (0.8580 - 0.9789)

Supplementary Table 12. RCC subtyping: ensemble performance evaluated on the BWH RCC independent test sets. For each slide, the predicted normalized scores from all 10 CLAM models developed on the TCGA training sets are averaged and used to inform the slide-level diagnosis. The ensemble performance is reported in terms of the macro-averaged test AUC, mAP and F1 score and balanced accuracy score (bACC). For individual subtypes, 95% confidence intervals for the per subtype one-vs-rest AUC were calculated using Delong's method and indicated in parentheses.

BWH Independent Test Set (n=131)				
Training Set Size	**AUC**	**mAP**	**F1**	**bACC**
100% of Train Data	0.9797 (0.9543 - 1.0000)	0.9859	0.6667	0.7500
75% of Train Data	0.9792 (0.9544 - 1.0000)	0.9846	0.6923	0.7647
50% of Train Data	0.9725 (0.9435 - 1.0000)	0.9779	0.7170	0.7794
25% of Train Data	0.9360 (0.8972 - 0.9749)	0.9508	0.6667	0.7500
10% of Train Data	0.9043 (0.8528 - 0.9558)	0.9163	0.4545	0.6471
BWH Independent Test Set (Cellphone, n=131)				
100% of Train Data	0.8826 (0.8241 - 0.9411)	0.8826	0.8188	0.7898
75% of Train Data	0.8695 (0.8088 - 0.9303)	0.8740	0.7742	0.7894
50% of Train Data	0.8527 (0.7872 - 0.9182)	0.8842	0.5474	0.6832
25% of Train Data	0.8140 (0.7395 - 0.8885)	0.8306	0.3333	0.5871
10% of Train Data	0.8583 (0.7922 - 0.9244)	0.8240	0.3133	0.5797
BWH Independent Test Set (Biopsy, n=110)				
100% of Train Data	0.9233 (0.8735 - 0.9731)	0.9308	0.4507	0.6455
75% of Train Data	0.9078 (0.8522 - 0.9633)	0.9256	0.4722	0.6545
50% of Train Data	0.9071 (0.8491 - 0.9651)	0.9274	0.4058	0.6273
25% of Train Data	0.8800 (0.8120 - 0.9480)	0.9079	0.2812	0.5818
10% of Train Data	0.8317 (0.7560 - 0.9074)	0.8565	0.0702	0.5182

Supplementary Table 13. NSCLC subtyping: ensemble performance evaluated on the BWH NSCLC independent test sets. For each slide, the predicted normalized scores from all 10 CLAM models developed on the TCGA + CPTAC training sets are averaged and used to inform the slide-level diagnosis. The ensemble performance is reported in terms of the average test AUC, mAP, F1 score and balanced accuracy score (bACC). The 95% confidence intervals for the true AUC were calculated using Delong's method and indicated in parentheses.

BWH Independent Test Set (n=133)				
Training Set Size	**AUC**	**mAP**	**F1**	**bACC**
100% of Train Data	0.9491 (0.9133 - 0.9850)	0.9597	0.8889	0.8872
75% of Train Data	0.9123 (0.8629 - 0.9616)	0.9009	0.8201	0.8120
50% of Train Data	0.8856 (0.8267 - 0.9444)	0.9076	0.8116	0.8045
25% of Train Data	0.8252 (0.7538 - 0.8966)	0.8607	0.7317	0.7519
10% of Train Data	0.8089 (0.7331 - 0.8847)	0.8442	0.7419	0.7594

Supplementary Table 14. Lymph node metastasis detection: ensemble performance evaluated on the BWH lymph node independent test sets. For each slide, the predicted normalized scores from all 10 CLAM models developed on the Camelyon16 + Camelyon17 training sets are averaged and used to inform the slide-level diagnosis. The ensemble performance is reported in terms of the average test AUC, mAP, F1 score and balanced accuracy score (bACC). The 95% confidence intervals for the true AUC were calculated using Delong's method and indicated in parentheses.

Training Data	AUC	bACC	F1	mAP
100% of Train Data	0.9213 ± 0.0234	0.6181 ± 0.0817	0.6050 ± 0.0943	0.8670 ± 0.0350
75% of Train Data	0.9074 ± 0.0247	0.5564 ± 0.0921	0.5201 ± 0.1237	0.8424 ± 0.0474
50% of Train Data	0.9002 ± 0.0258	0.6299 ± 0.0855	0.5996 ± 0.1131	0.8391 ± 0.0447
25% of Train Data	0.9265 ± 0.0146	0.7065 ± 0.0617	0.6900 ± 0.0751	0.8865 ± 0.0238
10% of Train Data	0.9097 ± 0.0217	0.7037 ± 0.0858	0.6998 ± 0.0938	0.8571 ± 0.0335

Supplementary Table 15. RCC subtyping performance evaluated on the BWH cellphone microscopy image test set. The 10-fold average performance (± std) of CLAM models trained on TCGA are reported in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) for n = 135. Macro-averaging is used for one-vs-rest AUC, F1 and mAP.

Training Data	AUC	bACC	F1	mAP
100% of Train Data	0.8729 ± 0.0246	0.7716 ± 0.0418	0.7973 ± 0.0351	0.8714 ± 0.0311
75% of Train Data	0.8496 ± 0.0413	0.7431 ± 0.0621	0.7001 ± 0.1211	0.8512 ± 0.0541
50% of Train Data	0.8368 ± 0.0391	0.6654 ± 0.0918	0.4932 ± 0.2492	0.8640 ± 0.0298
25% of Train Data	0.7980 ± 0.0193	0.6078 ± 0.0640	0.3603 ± 0.1810	0.8188 ± 0.0159
10% of Train Data	0.8258 ± 0.0489	0.6233 ± 0.0785	0.4145 ± 0.2399	0.8055 ± 0.0436

Supplementary Table 16. NSCLC subtyping performance evaluated on the BWH cellphone microscopy image test set. The 10-fold average performance (± std) of CLAM models trained on TCGA + CPTAC are reported in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) for n = 131.

NSCLC Slide ID	Biopsy Specimens Embedded	Label	RCC Slide ID	Biopsy Specimens Embedded	Label
Slide_1	2	LUAD	Slide_1	3	CCRCC
Slide_2	2	LUAD	Slide_2	3	CCRCC
Slide_3	2	LUAD	Slide_3	3	CCRCC
Slide_4	1	LUAD	Slide_4	3	CCRCC
Slide_5	1	LUAD	Slide_5	3	CCRCC
Slide_6	1	LUAD	Slide_6	3	CCRCC
Slide_7	1	LUAD	Slide_7	2	CCRCC
Slide_8	1	LUAD	Slide_8	2	CCRCC
Slide_9	2	LUAD	Slide_9	3	CCRCC
Slide_10	3	LUAD	Slide_10	3	CCRCC
Slide_11	2	LUAD	Slide_11	2	CCRCC
Slide_12	2	LUAD	Slide_12	2	CCRCC
Slide_13	2	LUAD	Slide_13	2	CCRCC
Slide_14	2	LUAD	Slide_14	1	CCRCC
Slide_15	1	LUAD	Slide_15	1	CCRCC
Slide_16	2	LUAD	Slide_16	2	CCRCC
Slide_17	3	LUAD	Slide_17	2	CCRCC
Slide_18	2	LUAD	Slide_18	1	CCRCC
Slide_19	1	LUAD	Slide_19	1	CCRCC
Slide_20	2	LUAD	Slide_20	1	CCRCC
Slide_21	1	LUAD	Slide_21	1	CCRCC
Slide_22	1	LUAD	Slide_22	4	CCRCC
Slide_23	1	LUAD	Slide_23	4	CCRCC
Slide_24	2	LUAD	Slide_24	4	CCRCC
Slide_25	2	LUAD	Slide_25	2	CCRCC
Slide_26	2	LUAD	Slide_26	2	CCRCC
Slide_27	1	LUAD	Slide_27	2	CCRCC
Slide_28	1	LUAD	Slide_28	2	CCRCC
Slide_29	1	LUAD	Slide_29	2	CCRCC
Slide_30	4	LUAD	Slide_30	2	CCRCC
Slide_31	2	LUAD	Slide_31	2	CCRCC
Slide_32	2	LUAD	Slide_32	2	CCRCC
Slide_33	2	LUAD	Slide_33	2	CCRCC
Slide_34	2	LUAD	Slide_34	2	CCRCC
Slide_35	2	LUAD	Slide_35	3	CCRCC
Slide_36	2	LUAD	Slide_36	3	CCRCC
Slide_37	3	LUAD	Slide_37	2	CCRCC
Slide_38	1	LUAD	Slide_38	2	CCRCC
Slide_39	1	LUAD	Slide_39	2	CCRCC
Slide_40	5	LUAD	Slide_40	2	CCRCC
Slide_41	2	LUAD	Slide_41	4	CCRCC
Slide_42	3	LUAD	Slide_42	2	CCRCC
Slide_43	2	LUAD	Slide_43	2	CCRCC
Slide_44	2	LUAD	Slide_44	4	CCRCC
Slide_45	2	LUAD	Slide_45	4	CCRCC
Slide_46	5	LUAD	Slide_46	2	CCRCC
Slide_47	2	LUAD	Slide_47	1	CCRCC
Slide_48	1	LUAD	Slide_48	2	CCRCC
Slide_49	3	LUAD	Slide_49	2	CCRCC
Slide_50	3	LUAD	Slide_50	5	CCRCC
Slide_51	3	LUAD	Slide_51	5	CCRCC
Slide_52	3	LUAD	Slide_52	4	CCRCC
Slide_53	2	LUAD	Slide_53	4	CCRCC
Slide_54	4	LUAD	Slide_54	3	CRCC
Slide_55	2	LUAD	Slide_55	3	CRCC
Slide_56	1	LUSC	Slide_56	3	CRCC
Slide_57	4	LUSC	Slide_57	3	CRCC

Slide_58	5	LUSC	Slide_58	3	CRCC
Slide_59	2	LUSC	Slide_59	3	CRCC
Slide_60	4	LUSC	Slide_60	2	CRCC
Slide_61	4	LUSC	Slide_61	2	CRCC
Slide_62	6	LUSC	Slide_62	3	CRCC
Slide_63	2	LUSC	Slide_63	1	CRCC
Slide_64	4	LUSC	Slide_64	3	CRCC
Slide_65	5	LUSC	Slide_65	1	CRCC
Slide_66	2	LUSC	Slide_66	4	CRCC
Slide_67	3	LUSC	Slide_67	1	PRCC
Slide_68	4	LUSC	Slide_68	1	PRCC
Slide_69	5	LUSC	Slide_69	2	PRCC
Slide_70	2	LUSC	Slide_70	2	PRCC
Slide_71	4	LUSC	Slide_71	2	PRCC
Slide_72	2	LUSC	Slide_72	2	PRCC
Slide_73	3	LUSC	Slide_73	2	PRCC
Slide_74	4	LUSC	Slide_74	3	PRCC
Slide_75	2	LUSC	Slide_75	2	PRCC
Slide_76	2	LUSC	Slide_76	2	PRCC
Slide_77	3	LUSC	Slide_77	3	PRCC
Slide_78	5	LUSC	Slide_78	3	PRCC
Slide_79	4	LUSC	Slide_79	2	PRCC
Slide_80	3	LUSC	Slide_80	2	PRCC
Slide_81	3	LUSC	Slide_81	2	PRCC
Slide_82	3	LUSC	Slide_82	1	PRCC
Slide_83	4	LUSC	Slide_83	1	PRCC
Slide_84	3	LUSC	Slide_84	1	PRCC
Slide_85	1	LUSC	Slide_85	1	PRCC
Slide_86	2	LUSC	Slide_86	1	PRCC
Slide_87	1	LUSC	Slide_87	1	PRCC
Slide_88	4	LUSC	Slide_88	4	PRCC
Slide_89	2	LUSC	Slide_89	4	PRCC
Slide_90	1	LUSC	Slide_90	2	PRCC
Slide_91	1	LUSC	Slide_91	2	PRCC
Slide_92	4	LUSC	Slide_92	1	PRCC
Slide_93	3	LUSC			
Slide_94	3	LUSC			
Slide_95	1	LUSC			
Slide_96	4	LUSC			
Slide_97	2	LUSC			
Slide_98	5	LUSC			
Slide_99	1	LUSC			
Slide_100	2	LUSC			
Slide_101	2	LUSC			
Slide_102	2	LUSC			
Slide_103	2	LUSC			
Slide_104	1	LUSC			
Slide_105	4	LUSC			
Slide_106	2	LUSC			
Slide_107	1	LUSC			
Slide_108	1	LUSC			
Slide_109	4	LUSC			
Slide_110	1	LUSC			

Supplementary Table 17. Number of biopsy specimens embedded on BWH In-house biopsy slides. The number of biopsy specimens embedded on each slide varies and ranges from 1 – 6 for lung biopsy WSIs and 1 - 5 for kidney biopsy WSIs. The median number of embedded specimens per slide is 2 for both datasets.

Training Data	AUC	bACC	F1	mAP
100% of Train Data	0.9514 ± 0.0110	0.7983 ± 0.0341	0.7164 ± 0.0519	0.9125 ± 0.0204
75% of Train Data	0.9318 ± 0.0150	0.7455 ± 0.0378	0.6573 ± 0.0580	0.8925 ± 0.0269
50% of Train Data	0.9206 ± 0.0126	0.7067 ± 0.0659	0.6470 ± 0.0671	0.8358 ± 0.0267
25% of Train Data	0.9346 ± 0.0191	0.7452 ± 0.0325	0.7350 ± 0.0365	0.8649 ± 0.0356
10% of Train Data	0.9232 ± 0.0260	0.6960 ± 0.0641	0.7150 ± 0.0693	0.8508 ± 0.0567

Supplementary Table 18. RCC subtyping performance evaluated on the BWH RCC biopsy test set. The 10-fold average performance (± std) of CLAM models trained on TCGA are reported in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) for n = 92.

Training Data	AUC	bACC	F1	mAP
100% of Train Data	0.9017 ± 0.0161	0.6509 ± 0.0455	0.4593 ± 0.1123	0.9133 ± 0.0139
75% of Train Data	0.8819 ± 0.0250	0.6555 ± 0.0367	0.4731 ± 0.0930	0.9009 ± 0.0238
50% of Train Data	0.8848 ± 0.0106	0.6309 ± 0.0371	0.4108 ± 0.0885	0.9075 ± 0.0089
25% of Train Data	0.8457 ± 0.0256	0.5973 ± 0.0390	0.3187 ± 0.1082	0.8747 ± 0.0248
10% of Train Data	0.7881 ± 0.0742	0.5282 ± 0.0131	0.1393 ± 0.0903	0.8023 ± 0.0842

Supplementary Table 19. NSCLC subtyping performance evaluated on the BWH NSCLC biopsy test set. The 10-fold average performance (± std) of CLAM models trained on TCGA + CPTAC are reported in terms of test AUC, mean average precision score (mAP), F1 score and balanced accuracy score (bACC) for n = 110.

Dataset	Link
TCGA Kidney RCC	https://portal.gdc.cancer.gov/repository
TCGA Lung NSCLC	https://portal.gdc.cancer.gov/repository
CPTAC Lung NSCLC	https://cancerimagingarchive.net/datascope/cptac/
Camelyon16 + Camelyon17	https://camelyon17.grand-challenge.org/

Supplementary Table 20. Access links to public datasets used.