Interpretable Tumor Localization in Bladder Cancer Histopathology Using Deep Multiple Instance Learning

by

Karthik Nair

SB, Computer Science and Molecular Biology, Massachusetts Institute of Technology, 2022

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Computer Science and Molecular Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2022

Author . . . . . .

Department of Electrical Engineering and Computer Science

May 13, 2022

Certified By . . . . . .

Eliezer Van Allen

Associate Professor

Thesis Supervisor

Certified By . . . . . .

Caroline Uhler

Associate Professor

Thesis Supervisor

Accepted By . . . . . .

Katrina LaCurts

Chair, Master of Engineering Thesis Committee

Interpretable Tumor Localization in Bladder Cancer Histopathology Using Deep Multiple Instance Learning

by

Karthik Nair

Submitted to the Department of Electrical Engineering and Computer Science on May 13, 2022 in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Computer Science and Molecular Biology

Abstract

Deep learning has emerged in cancer histopathology as a tool for predicting clinical and molecular properties of a patient's disease, thereby connecting slide with function. This concept is especially relevant to bladder cancer, where molecular and histopathologic heterogeneity is known to impact oncogenesis and disease progression, though the underlying properties governing these features are incompletely known. Traditional tile-based deep learning approaches to analyze bladder cancer (and other cancer) histopathology images do not integrate information across whole slides, potentially forfeiting accuracy and interpretability on more complex pathology tasks that require global slide context beyond local morphology, such as tumor subtyping and mutation prediction. To this end, we compare CLAM, a recently developed multiple instance learning model designed to address these limitations, to a tile-based computer vision model on over 1,500 hematoxylin and eosin (H&E)-stained bladder cancer (urothelial carcinoma) slides. We found that CLAM was more robust against overfitting to spurious confounders when compared with a traditional approach, resulting in more interpretable outputs. Additionally, we generated high-resolution tumor localization maps for a previously unstudied cohort using CLAM. Taken together, our results demonstrate CLAM to be a promising approach for tackling difficult digital pathology tasks previously hindered by traditional approaches.

Thesis Supervisors:

Dr. Eliezer Van Allen, Associate Professor of Medicine at Harvard Medical School
Dr. Caroline Uhler, Associate Professor of EECS at MIT

Acknowledgements

I've had an amazing time working in the Van Allen lab these past 9 months, from seeing the incredible science being done to random conversations over lunches and snack breaks. In the process, I've realized how important it is to foster a cooperative team that connects both on and beyond the work being done. Thus, I owe a massive thanks to my PI, Eliezer Van Allen, for bringing me into such a supportive group, providing frequent feedback through our regular meetings, and navigating the MEng process with me for the first time. On this last point, I would also like to thank Dr. Caroline Uhler for being my MEng co-advisor and making my collaboration with the Van Allen lab so seamless.

The image machine learning team I've been a part of here has been extremely welcoming. First, I owe thanks to Surya for being an incredible mentor and providing constant, honest feedback on my project. He is no doubt a driving personality in our lab, never sparing us from a good gradient descent joke. Additionally, I owe thanks to Jackson for being one of the reasons I joined the lab in the first place. As yet another driving personality, he brings charisma and experience to our team as one of the senior members. Finally, I am deeply appreciative for the other members of the team during my time here - Nicita, Bowen, Pasha, Jingxin - for great discussions and advice that have pushed the work in this thesis forward.

Finally, I want to thank my close friends and family for leading my journey here and for supporting me always!

Table of Contents

Introduction

Cancer is a disease of aberrant growth of cells, accounting for 18 million medical diagnoses in 2020 worldwide and being a leading cause of death[1]. Diagnosis and prognosis of an individual's cancer are performed after biopsies of the tumor are obtained, processed, and examined under a microscope by pathologists. The morphological features and spatial organization of tissue upon staining by haematoxylin and eosin (H&E) reveal the type and stage of cancer, and H&E remains the most widely used stain for medical diagnosis by pathologists[2] (Fig. 1a, b).

Due to the large sample sizes and rich readouts of H&E stains, deep learning has been used to extract complex features from these images relating to morphology and localization and subsequently predict phenotypes of interest, such as tumor grade, subtype, and mutation status[3–8]. Prediction of these phenotypes via deep learning allows for achieving or outperforming human performance on some tasks at scale[8], dissecting a slide for the regions most contributing to a phenotype[4], and stratifying patients by more complex phenotypes (e.g. immune infiltration[9]) to understand their relationship with disease progression.

Urothelial carcinoma in particular is a debilitating cancer for which diagnostics, treatment, and survival rates have been largely unchanged in three decades[10], owing to the extraordinary genomic and spatial heterogeneity of the bladder[11,12]. While 80% of patients are diagnosed with early stage disease and generally have good survival outcomes, the remaining 20% of patients have substantially poorer prognosis[10]. Resistance to chemotherapy, the standard-of-care option for these patients, is a major driver of poor outcomes for this later-stage class of patients[11]. Given the morbidity and uncertain clinical utility associated with these therapies, accurately predicting which patients are most likely to benefit from these treatments and the salient properties of their tumors would have significant clinical impact. Examples of outcomes-relevant predictive properties in bladder cancer include the presence of critical mutations[13] and tumor/transcriptional subtype[11].

These properties have been shown to be predictable using deep learning from whole-slide images (WSI) routinely collected for diagnosis without the need for time and cost-intensive additional experiments[6,7]. Yet, previous deep learning approaches often perform suboptimally on these complex tasks because slides are divided into smaller tiles during training to fit within memory and computing limits. This effectively distributes the slide-level label to potentially misrepresentative tiles, such as a tumor slide label being assigned to a stromal tile present far from the tumor region. The division of slides into tiles additionally obscures the larger

slide context from the model during training. Consequently, accuracy and interpretability of models on these clinically relevant tasks may be suboptimal.

However, a recent multiple instance learning framework for histopathology (CLAM) addresses this challenge by integrating local tile information across the slide before arriving at a slide-level prediction[14]. The authors show that CLAM is capable of producing more accurate and more interpretable predictions than tile-based methods. In particular, they show that CLAM is capable of extrapolating the slide-level label to pixel-level attention maps for each slide according to the regions of the slide that drive the model's prediction. Thus, we hypothesized that the CLAM framework could be used to improve prediction on the aforementioned tasks relevant to bladder cancer.

To this end, we investigated the utility of CLAM as compared to a tile-based method in predicting tumor/normal status, evaluating accuracy and interpretability of both approaches on over 1,500 urothelial carcinoma WSI's from two cohorts.

Chapter 1. Data Overview and Preprocessing

In this brief chapter, we perform exploratory analysis and preprocessing on the TCGA bladder cancer (TCGA-BLCA) and Dana Farber Cancer Institute PROFILE(DFCI) whole-slide image datasets.

These datasets are composed of fresh-frozen (FF) and formalin-fixed paraffin-embedded (FFPE) slides, which are two modalities for preserving tissue. While FF preparation better preserves DNA molecules by avoiding the harsh fixation and archiving process in FFPE preparation, FFPE slides better preserve cellular and architectural morphology of the tissue (Fig. 1)[15]. Additionally, FFPE slides can be archived and are more widespread than FF slides, explaining their high representation in these datasets.

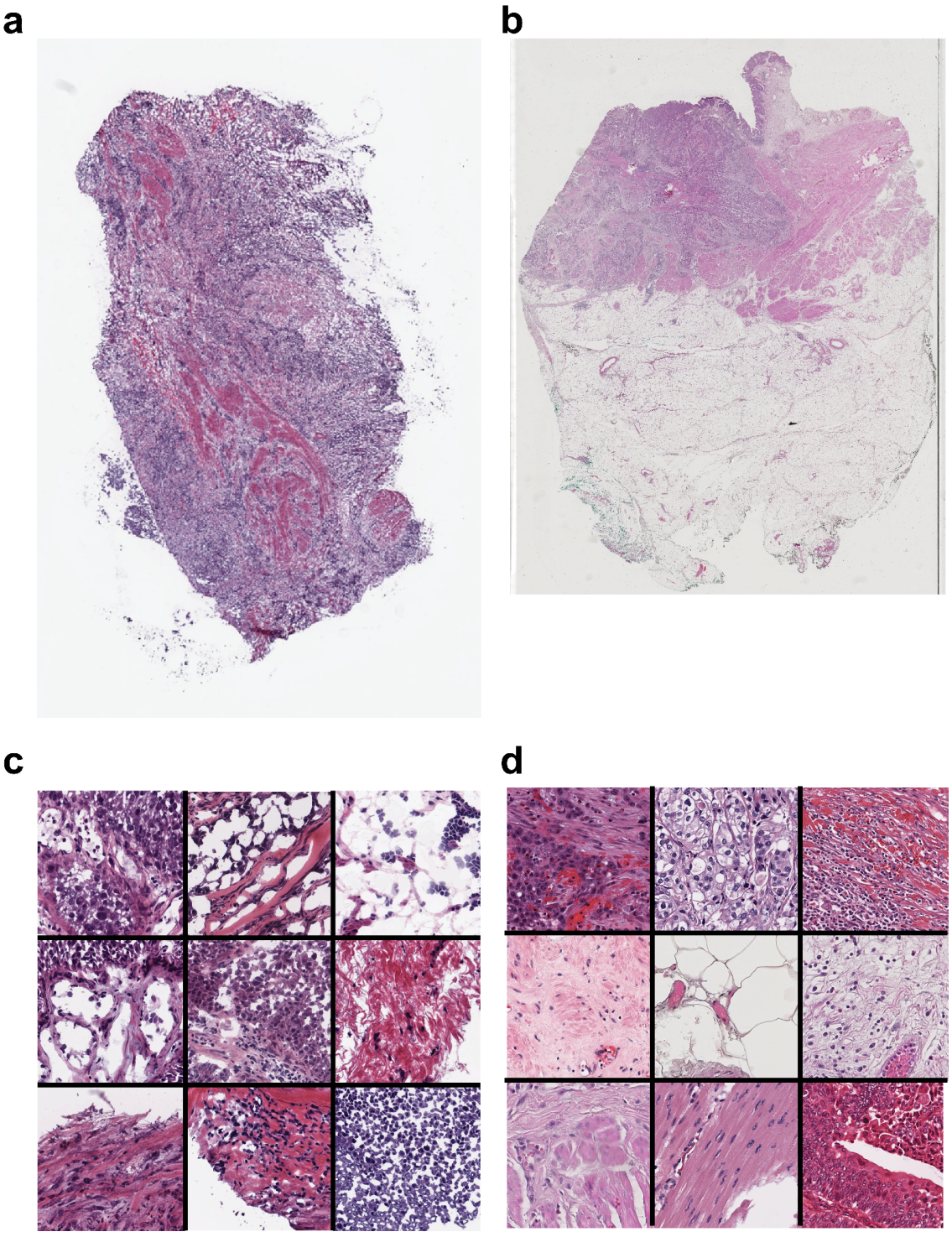

Figure 1. Examples of H&E-Stained Slides. All images are derived from the TCGA dataset. (a) Fresh-frozen (FF) and (b) fixed-formalin paraffin-embedded (FFPE) WSI's, the two prominent modalities for saving tissue specimens. 20x snapshots of (c) FF and (d) FFPE tiles are also shown.

1.1 Obtaining Data and Exploratory Analysis

TCGA-BLCA slides and patient data were obtained from the GDC Data Portal [16]. The dataset is composed of 469 fresh-frozen (FF) and 457 formalin-fixed paraffin-embedded (FFPE) slides from patients with urothelial carcinoma, with each slide represented as a gigapixel-resolution image (Fig. 2). To serve as an external validation dataset, the Dana Farber Cancer Institute Profile slides and clinical data were also obtained. This dataset is composed of 538 FFPE slides collected after diagnosis of urothelial carcinoma (Fig. 2).

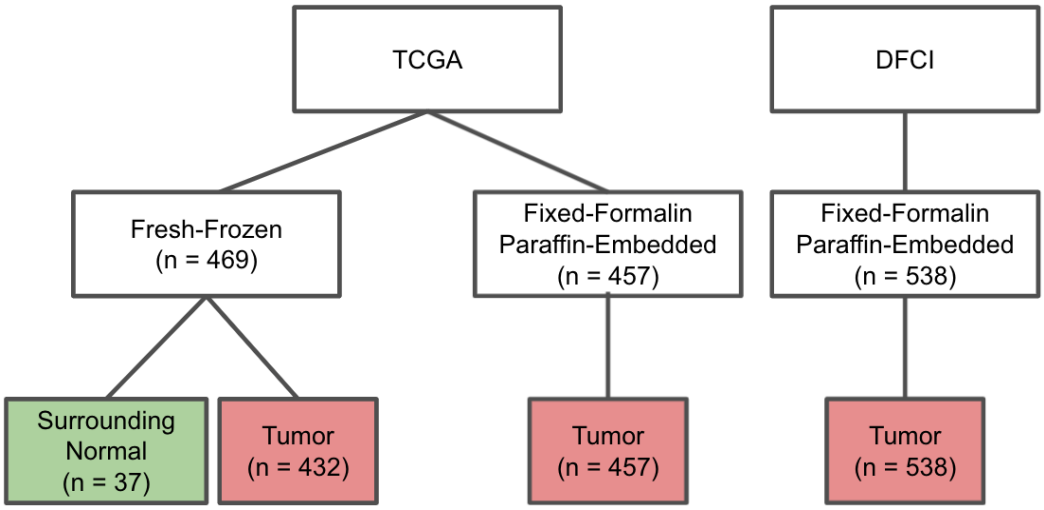

Figure 2. Histopathological Datasets Used in this Study.

The TCGA-BLCA dataset represents patients with bladder tumors across a spectrum of disease states and contains annotations for grade, stage, and tumor subtype[11]. Exploratory data analysis was performed to understand the covariation of these annotations, which could affect class-balancing and interpretation of downstream results (Fig. 3). A disproportionate number of cases were high-grade (94.2%) and late-stage (99.0% stage II and later), owing to the selection criteria of the TCGA-BLCA cohort. As expected, non-papillary subtyping, late stage, and high grade frequently co-occurred. Additionally, the luminal papillary subtype was enriched in less aggressive cases, while the other subtypes were generally enriched in later stage cases. These findings highlight previously-known properties of aggressive urothelial carcinoma in the TCGA setting[11].

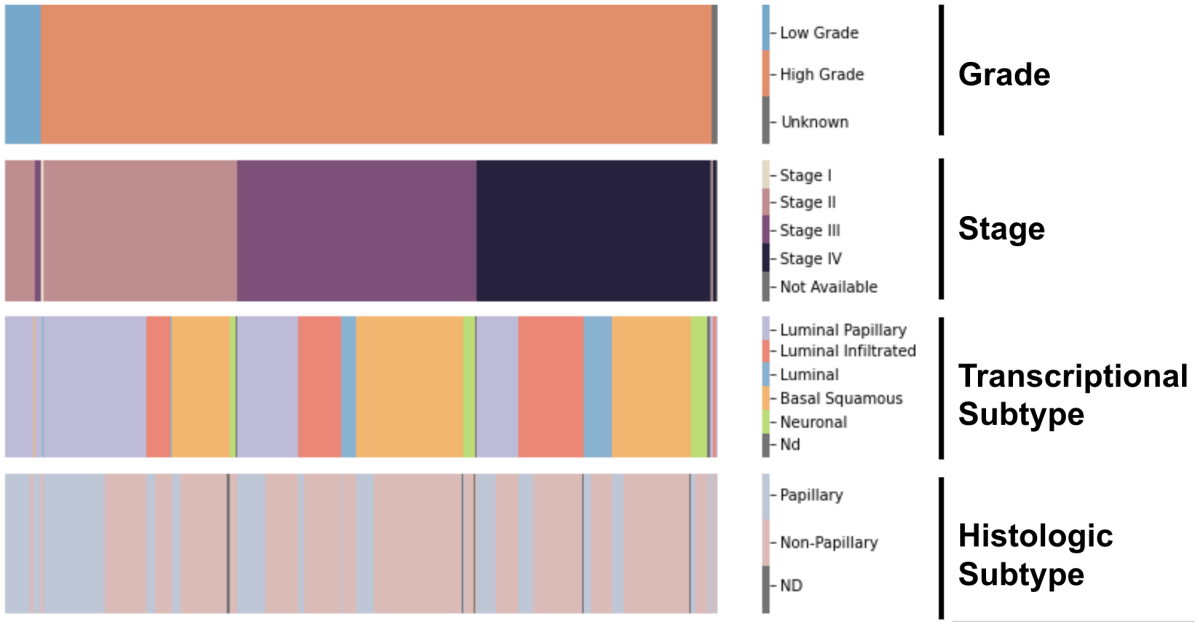

Figure 3. Exploratory Analysis of TCGA Dataset. Each patient is a column.

1.2 HistoQC Masking

Whole-slide images often contain artifacts that present confounders for downstream analyses, such as pen marks, whitespace, and blurred regions[17]. To correct for this, we used the *HistoQC* library to generate a mask of usable tissue region for each slide[18]. *HistoQC* is a histopathology quality control tool that filters out aberrant parts of the slide through a user-defined sequence of modules, incorporating kernel-based approaches as well as deep learning classifiers. We applied *HistoQC* to all FF and FFPE slides (Fig. 4). While the default configuration performs decently, we developed a further-tuned configuration that included more tissue regions. Notably, the resulting masks reliably excluded whitespace and pen marks. The further-tuned configuration also selected tissue regions with a higher whitespace fraction, such as adipose tissue, which may be relevant for downstream prediction tasks.

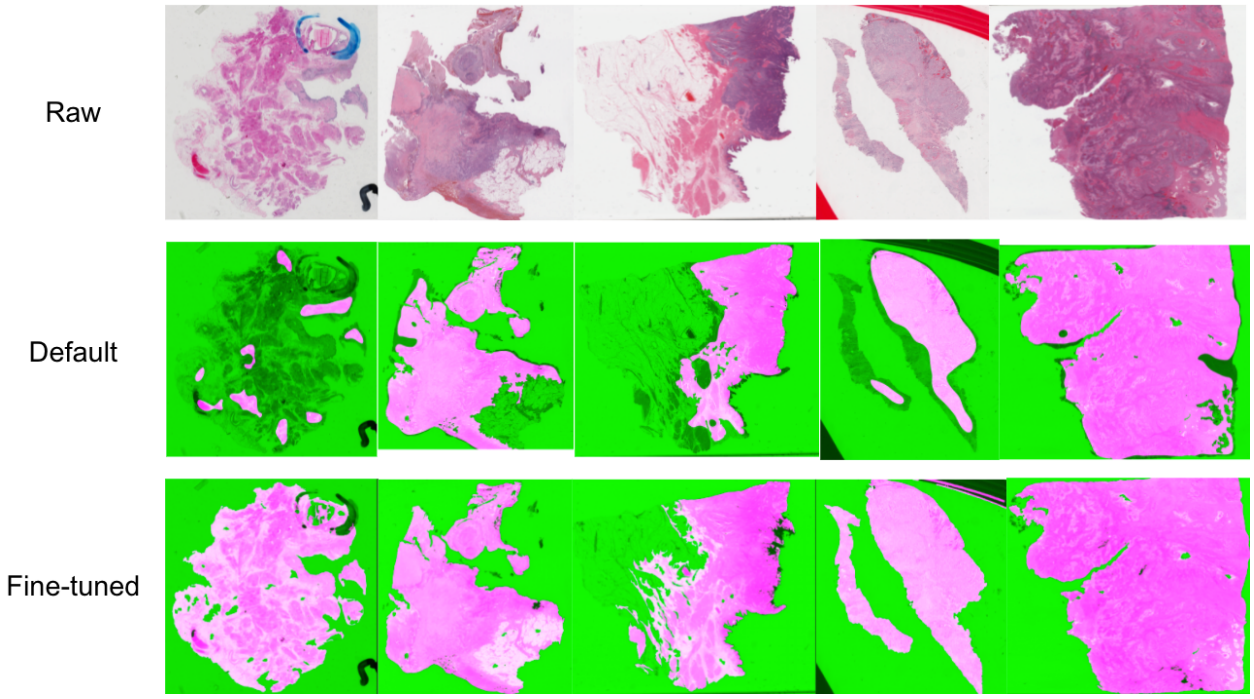

Figure 4. *HistoQC* Accurately Segments the Tissue Regions in FFPE Slides. The results for 5 FFPE slides are shown. Top row, input slide; middle row, *HistoQC* output using default configuration; bottom row, *HistoQC* output using tuned configuration. Pink, region included in mask; green, region excluded in mask.

1.3 Tiling

Armed with masks of tissue regions for each slide, we next performed tiling to subdivide each gigapixel-resolution WSI into smaller images capable of being passed through computer vision models. We tiled each WSI into 512 x 512 pixel pieces and retained tiles containing at least 75% of their area included in the mask. This resulted in a total of 209,956 FF tiles and 3,246,633 FFPE tiles across the TCGA and DFCI cohorts, with each slide harboring as few as 7 tiles or as many as 10,000 (Fig. 5). We observed that FF slides had substantially fewer tiles due to being smaller images than FFPE slides. Additionally, the TCGA FFPE slides yielded almost twice as many tiles as the DFCI slides. We hypothesize that this is due to increased whitespace and pen marks in the DFCI slides, leading to more exclusive masks and thus fewer tiles.

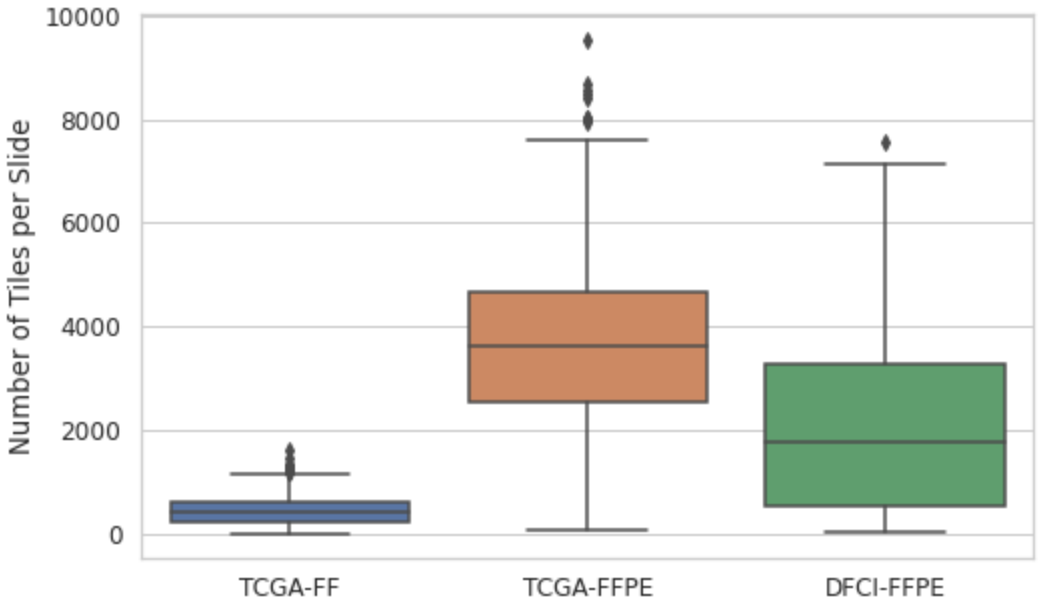

Figure 5. Cohort and Image Modality Differences Generate Different Tile Count Distributions. FF, fresh-frozen; FFPE, fixed-formalin paraffin-embedded.

Chapter 2. Tumor vs Normal Prediction

With WSI's represented as quality-controlled collections of tiles, in this chapter we predicted whether a patient's WSI contains tumor tissue (hereafter referred to as tumor/normal status). In the process, we evaluated two pipelines for accuracy and interpretability and visualized model-detected patterns across an unstudied urothelial carcinoma cohort.

The tumor/normal status prediction problem is a foundational task in digital pathology that allows for benchmarking models, dissecting tumor regions of WSI's, and enabling deeper analyses of spatial organization of tumors. Model outputs can be visually validated due to the well-known morphology of tumors, allowing a deep understanding of model behavior. In particular, bladder tumor morphology can be characterized by nuclear crowding, irregular nuclear shape, frequent mitosis, and invasion of stroma and vasculature[8] (Fig. 6). Additionally, model outputs can be used to improve prediction of more challenging properties of tumors, as shown in prior work [6,8].

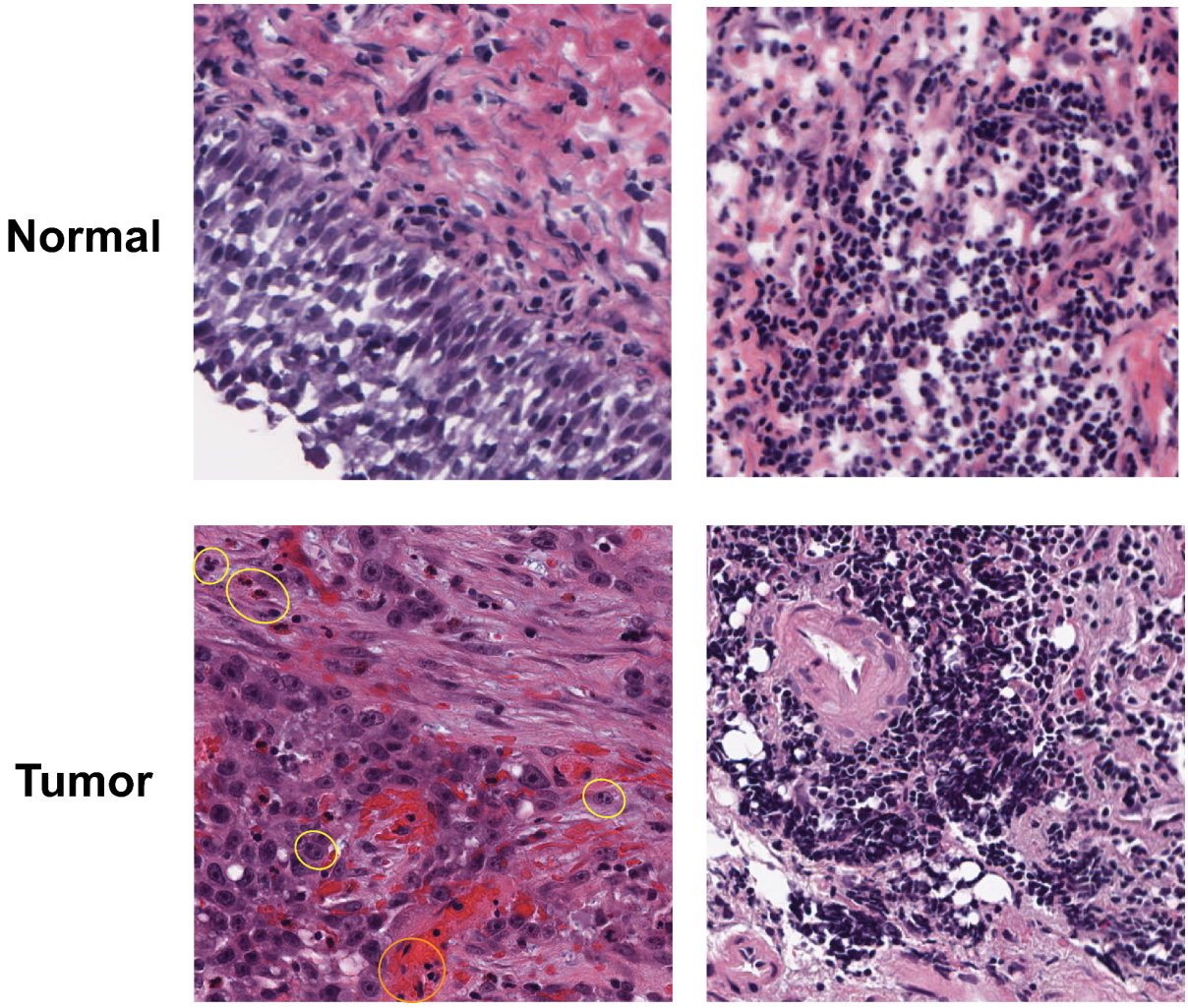

Figure 6. Examples of Tumor and Normal Tiles. Snapshots taken at 20x resolution. Yellow, examples of cells undergoing mitosis; orange, vascular infiltration.

Thus, we next trained deep neural networks to predict tumor/normal status. Our purpose of applying these models to this task is three-fold: (1) we can use the model to identify which features of a slide best explain its tumor or normal label, (2) we can subsequently segment the slide into tumor and stromal regions, and (3) we can embed tiles into vectors that can be used for other prediction tasks.

Methods

2.1 Data Setup

The tumor vs normal task uses the tiles generated in the previous step as input images to predict whether the tile represents tumor or normal tissue.

Since only slide-level labels of tumor vs normal are available for the TCGA and DFCI cohorts, we used a weakly-labeled scheme in which all tiles from a slide were assigned the slide-level label. However, this can result in misleading tile-label pairings in tumor-labeled slides due to the spatial heterogeneity of tumors across a WSI. For example, a tile sampled from a benign region of a tumor-labeled slide will be assigned the tumor label, despite presenting the normal phenotype. Thus, although weak-label effects may be negligible across entire slides, they place an upper limit on the accuracy we can expect, as well as complicate the learning of specific features from a slide.

Previous approaches to this task in bladder cancer include coping with the weakly-labeled strategy, and manually annotating tumor regions of interest for a strongly-labeled strategy [6,7]. For the tumor vs normal task, weak-labels have been shown to be sufficient to learn a good classifier[6]. Additionally, bladder cancer has a relatively high tumor-to-stroma ratio[6], suggesting that false positive tiles through weak-labeling are less common. Thus, we proceeded with the weak-label strategy.

2.2 Experiment Choices

The FF/FFPE distinction of slides within the TCGA training cohort presented a challenge for developing a model robust to confounders and guided our experiment setup. In particular, FFPE slides in this dataset are only taken for tumor samples, with the normal samples being exclusively in FF slides (Fig. 2). Additionally, since our DFCI external validation cohort is composed solely of FFPE slides, we required a model that could predict on FFPE slides. Thus, our model needed both FF and FFPE slides in the training set. Hence, our first line of experiments task the model with distinguishing between FF normal and FFPE tumor slides (henceforth termed the FF/FFPE experiment).

However, under this setup the model may learn to distinguish FF from FFPE rather than tumor vs normal, and this will be sufficient to train a well-performing yet misguided model. To address this, we pursued a second line of experiments using a transfer learning approach in

which the model learned to distinguish between FF normal and FF tumor samples (henceforth the FF/FF experiment). This FF-trained model could be then applied directly to predict on FFPE slides. Since the training set contains slides from only one imaging modality, this no longer presents a learnable confounder. Thus, we proceeded with the FF normal vs FF tumor experiment in addition to the aforementioned FF normal vs FFPE tumor experiment (Table 1).

Experiment	Normal Samples	Tumor Samples	Internally tested (TCGA) on	Externally validated (DFCI) on
FF/FFPE	FF	FFPE	FF normal and FFPE tumor	FFPE tumor
FF/FF	FF	FF	FF normal and FF tumor	FFPE tumor

Table 1. Machine Learning Experiment Choices.

2.3 Pipeline Choices

We tested two digital pathology workflows for prediction - 1) MC Lightning[19], a tile-based model using a pre-trained and fine-tuned ResNet[20] network, and 2) CLAM[14], an attention-based multiple instance learning model that integrates tile-level information to output slide level predictions and heatmaps (Fig. 7).

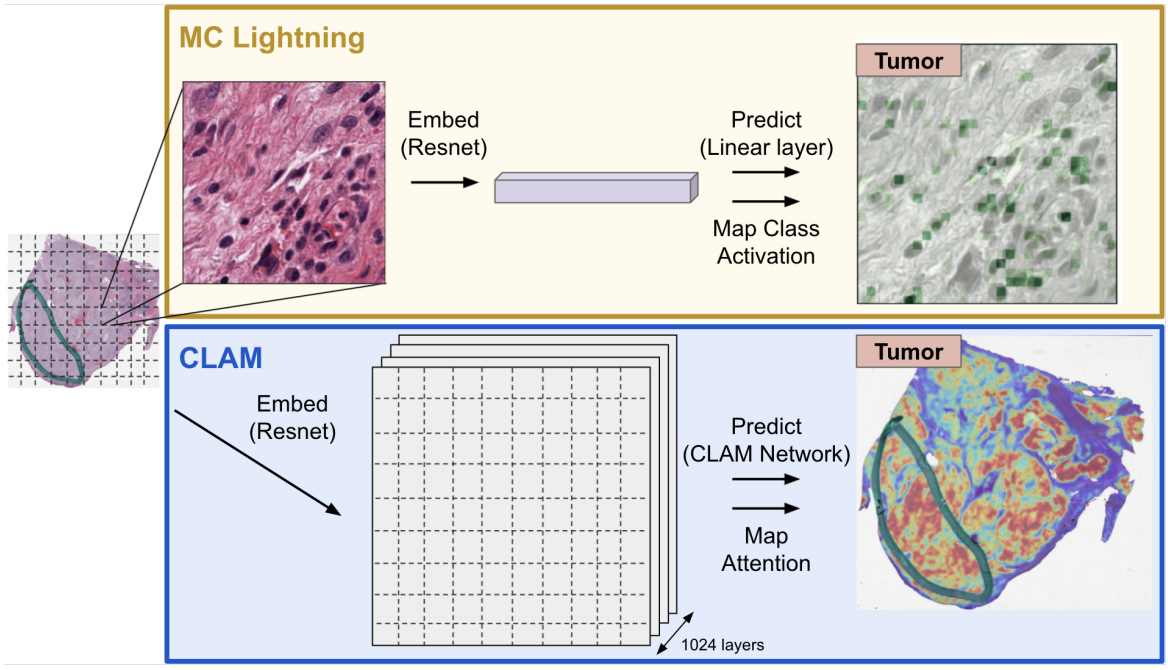

Figure 7. Tile-based MC Lightning and Slide-based CLAM Pipelines. Top panel heatmap - green corresponds to activated regions (per Grad-CAM[21]) for label prediction. Bottom panel heatmap - red represents high attention regions for tumor prediction, while blue represents low attention regions for tumor prediction.

2.3.1 MC Lightning Pipeline

MC Lightning is a tile-based deep learning pipeline developed internally for histopathology prediction tasks. Tile-based models take single tiles as inputs and predict a label for the single tile, independent of all other tiles in the slide.

We processed the 512 x 512 images resulting from the tiling procedure within each training epoch before using them as inputs to the model. Tiles were first randomly cropped to a 224 x 224 window and randomly flipped. This has been shown to increase training variability between epochs and consequently improve model generalizability[22]. Tiles were then RGB channel intensity-normalized to the ImageNet dataset for maximum compatibility when using the ImageNet-pretrained ResNet[20,23]. Resulting images were then used as inputs to the model. The model is composed first of a ResNet50 network that transforms images into an embedding. The output of the ResNet50 is a 2048-element embedding, which is passed through a dropout module that randomly zeros elements with a probability of 20%. We routed the dropped-out embedding directly into a linear layer that predicts the signal of each class in the image. Finally, we apply a sigmoid activation to generate the probability of each class.

The ResNet50 is a well-known neural network architecture that has served as the foundation for recent advances in computer vision, and we chose this architecture for its effectiveness in encoding image properties as vectors[20]. We unfroze the weights of the ResNet to allow for fine-tuning of the network to identify new image properties relevant specifically for the prediction task. Additionally, we use a linear layer to transform the embeddings to predictions for simplicity, generalizability, and interpretability.

2.3.2 CLAM Pipeline

CLAM is a slide-based deep learning pipeline for improving performance and interpretability in weakly-supervised histopathology tasks[14].

Similar to MC Lightning, an ImageNet-pretrained ResNet50 network is first used to embed each tile as a latent vector. In CLAM, however, the ResNet50 is frozen during training,

and subsequent attention layers convert the latent vector into an attention score encoding the model-assigned relevance of this tile for tumor/normal status prediction and used to weigh the overall slide prediction. CLAM thus learns important and unimportant regions of the slide for tumor/normal prediction and integrates the global context of the slide into prediction.

2.4 Data Splitting

We performed data splitting to evaluate accuracy of the trained model on unseen data (Table 2). Slides were first downsampled to 37 tumor and 37 normal slides so that the two classes were balanced. 15% of the slides were reserved for testing, and the remaining 85% were split into four folds for cross-validation. Each slide was subsampled to yield 400 tiles without replacement (or the number of tiles in the slide if less than 400). Both the MC Lightning and CLAM methods use the same splits of the data for comparable evaluation.

	Train	Validation	Test
FF/FFPE	14,980 (46.5)	4,993 (15.5)	4,208 (12)
FF/FF	14,020 (49)	3,534 (13)	3,988 (12)

Table 2. Tumor vs Normal Split Sizes. The number of tiles (averaged across 4 folds) for both experiments, with the corresponding average number of slides in parentheses. Notably, the FF/FF validation and test slides are all FF — the FF to FFPE transfer was not assessed in the TCGA split.

2.5 Model Training and Performance

Both MC Lightning and CLAM models were trained on TCGA training data for the tumor vs normal task. The MC Lightning models were trained for 10 epochs with a batch size of 32 tiles, while the CLAM models were trained for 30 epochs with a batch size of 1 slide. Metrics were logged using the Weights and Biases (W&B) software[24]. Learning rate and batch size were tuned using W&B's Bayesian search algorithm to minimize validation loss.

Results

At the end of model training, validation and test accuracy on the TCGA dataset were computed for the MC Lightning models across tiles for each of the 4 folds (Fig. 8). The model achieved surprisingly high tile-level accuracy, with median test accuracies of 98.2% (FF/FFPE) and 99.9% (FF/FF). When collapsing tile-level predictions into slide-level predictions via average pooling and evaluating on all slides not in the training set for each fold, the models achieve class-balanced accuracies of 99.9%, 98.0%, and 95.2% for MCL FF/FFPE, MCL FF/FF, and CLAM FF/FFPE respectively (Fig. 9). Thus, all three models generalize to unseen slides with high accuracy within the TCGA cohort.

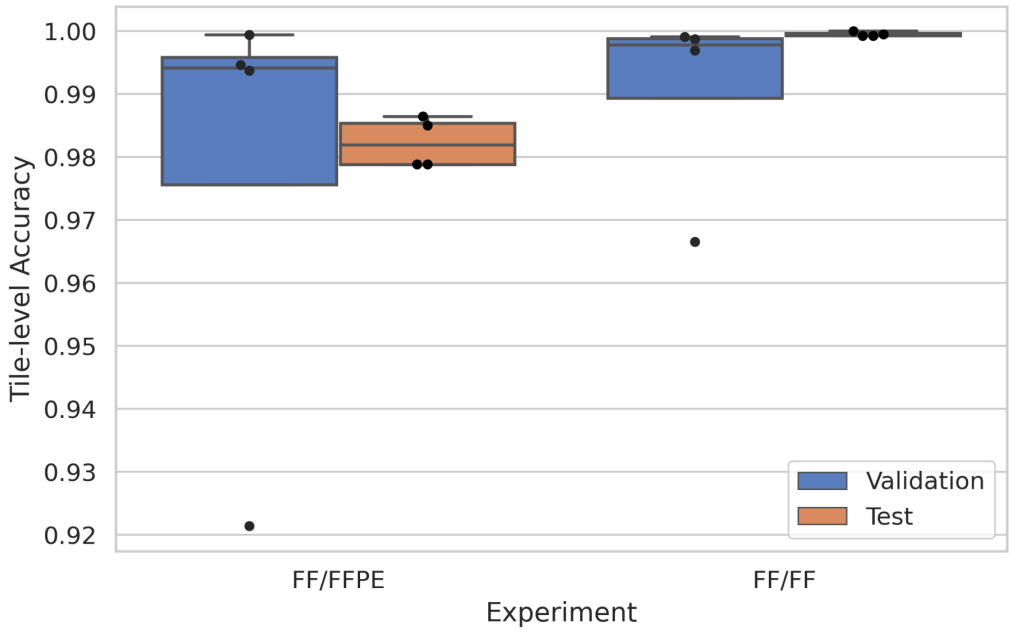

Figure 8. Tile-Level Accuracy of MC Lightning. Each dot represents the performance for a single training fold. FF, flash-frozen; FFPE, fixed-formalin paraffin-embedded. Experiments: FF/FFPE - distinguishing between FF normal and FFPE tumor slides; FF/FF - distinguishing between FF normal and FF tumor slides.

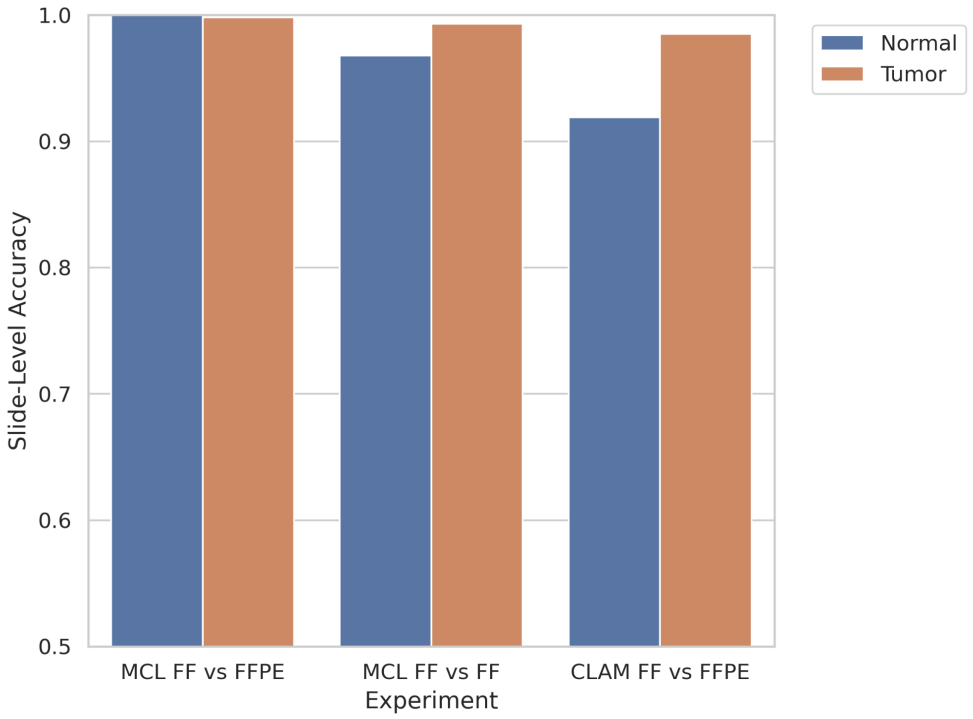

Figure 9. Slide-level Model Accuracies on TCGA Held-Out Data. Average held-out predictions for each slide were calculated by averaging predictions for all the fold models that did not contain the slide in the training set.

2.6 MC Lightning Models Learn Confounders

Although the high slide-level accuracy suggests the model is learning and generalizing to the test set, the tile-level accuracy of the MCL models is unreasonably high for this weakly-labeled task. Considering that many tiles from tumor slides will not harbor tumor-like morphology (e.g., tiles from non-infiltrated stroma), we expect a reasonable model to frequently mis-predict these tiles[25]. However, examining the predicted probabilities for stromal tiles from tumor slides versus normal slides suggests that there is information being learned from these tiles that polarizes the predictions toward the slide labels (Fig. 10). We explored two possible explanations for this: 1) train-test leakage and 2) confounding variables.

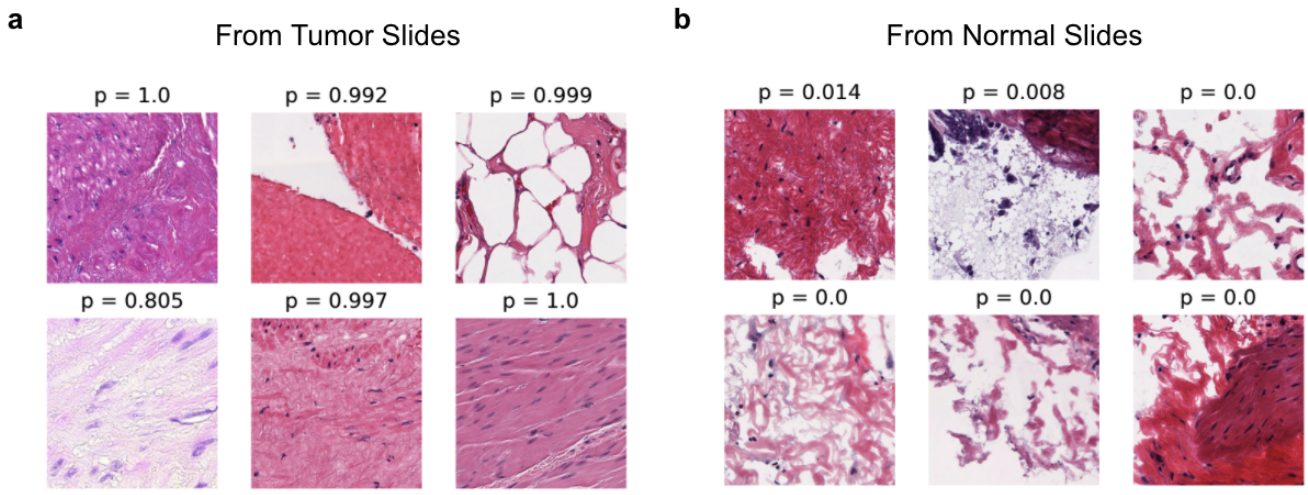

Figure 10. MC Lightning Model Overfits to Slide Label on Stromal Tiles. Selected stromal tiles from (a) FFPE tumor slides and (b) FF normal slides and their associated held-out predicted probabilities of tumor by the MCL FF/FFPE model. Although stromal tiles are expected to be predicted normal in tumor slides due to often bearing a normal phenotype, the MCL model mistakenly predicts them to be tumor. p, model-predicted probability of being tumor.

Leakage of data between the train and test sets could explain why the model appears to overfit at the tile-level. Train-test leakage could occur as either the same slide appearing in both splits, or the same exact tile appearing in both splits. To check both cases, for each test-set tile, we identified its nearest neighbor train-set tile in the embedding space (Fig. 11). We weighted the euclidean distance metric by the coefficients of the final linear layer for the nearest-neighbor algorithm. This allows the prediction-relevant components of the 2048-length embeddings to contribute most to nearest-neighbor determination. We found that although the nearest-neighbor train-set tiles were similar in appearance and morphology to their corresponding test-set tiles, they were not trivial transformations of each other and belonged to different slides. Thus, we rejected train-test leakage as a possibility.

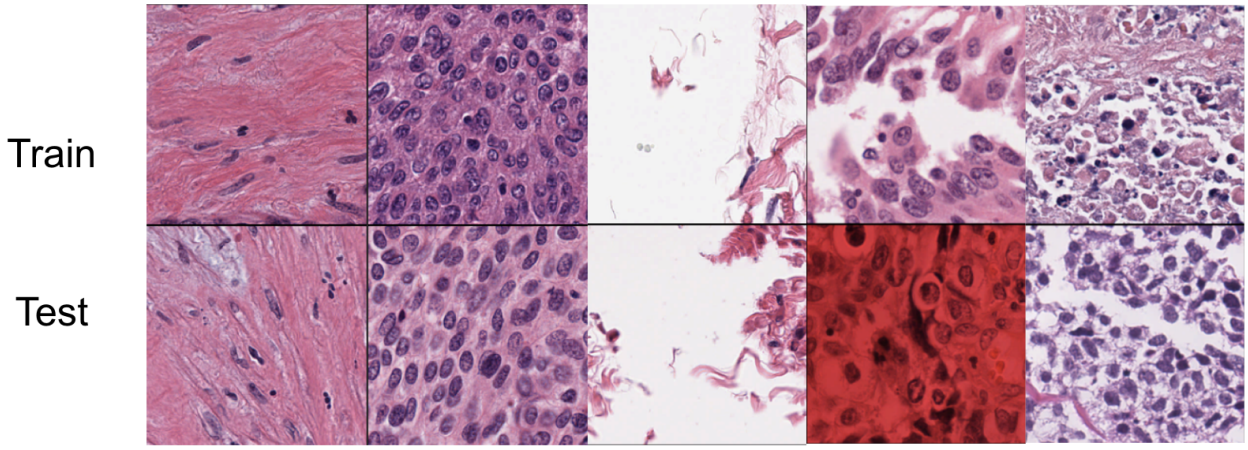

Figure 11. Nearest-neighbor tiles affirm no train-test leakage and quality of embeddings. Select test-set tiles and their corresponding nearest-neighbor train-set tiles according to ResNet50 embeddings.

Subtle confounding between the tumor and normal classes that present in most tiles could also explain the high tile-level accuracy by providing a consistent and learnable, yet artificial signal. The clearest confounder is the FF/FFPE status within the MCL FF/FFPE task, where learning how FF and FFPE tiles differ is sufficient to achieve perfect accuracy. However, the even higher tile-level accuracy of the MCL FF/FF model suggests that confounding may be present beyond this. To investigate the mechanism of confounding, we generated gradient-weighted class activation maps (Grad-CAM's) that pinpoint the regions of each tile that contribute most strongly to its prediction (Fig. 12)[21]. We found the model to attribute tumor predictions largely to nuclei, while attributing normal predictions to extracellular regions, a behavior consistent with the aberrant morphology and increased nuclear density of tumor tissue. However, whitespace was interestingly also attributed for normal prediction. When the whitespace forms a structure, e.g. a gland, the attribution may be capturing the lack of invading tumor cells (Fig. 12, row 3). But when the whitespace results from gaps or edges in the tissue that have no intrinsic biological meaning, it may represent the model learning a confounder (Fig. 12, row 2). Besides whitespace, H&E slides are known to contain many artifacts[17], and some of these may further explain the high tile-level accuracy of the MCL models, although these were not detected in the Grad-CAMs.

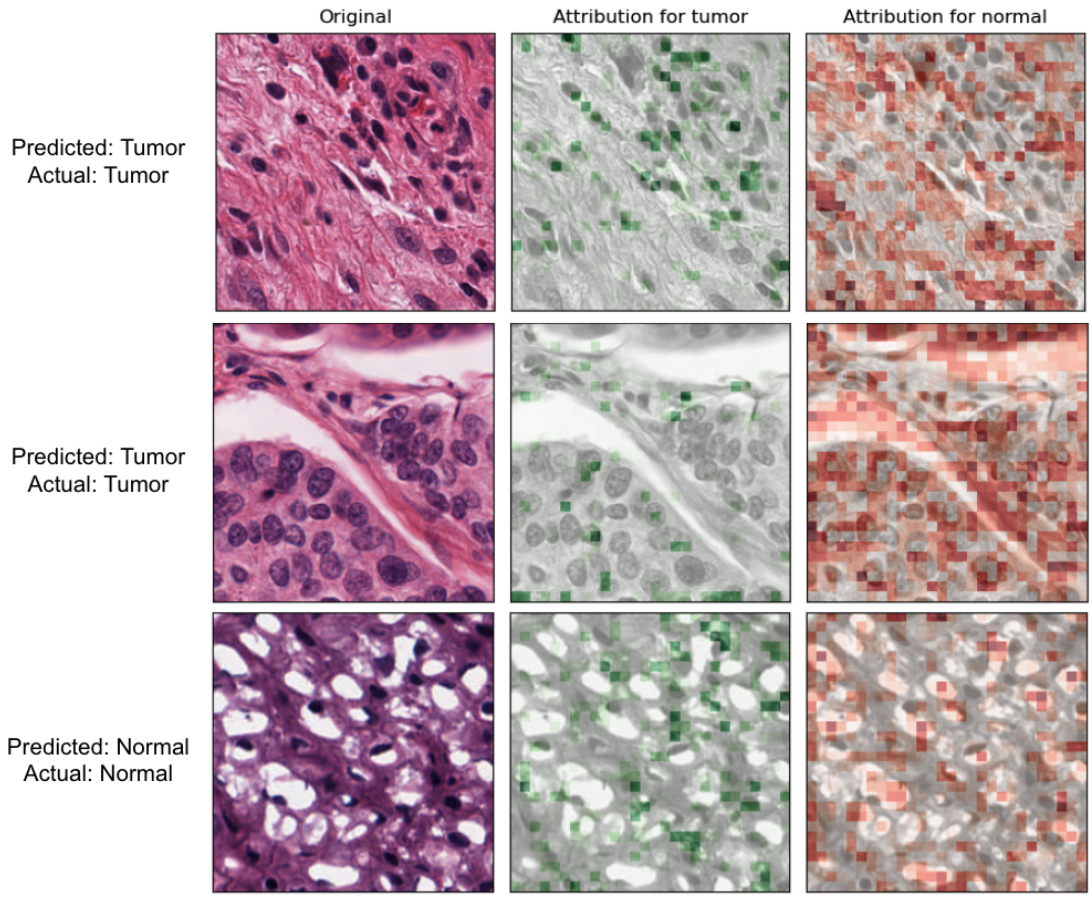

Figure 12. Grad-CAM Highlights Nuclei, Whitespace, and Extracellular Space as Important Features for MCL Tumor vs Normal Model.

2.7 External Validation on Dana Farber PROFILE Cohort

Faced with subtle confounders within the TCGA dataset that could provide a misleading assessment of performance, we next applied our TCGA-trained models to slides from the Dana Farber Cancer Institute (DFCI) PROFILE dataset to assess model generalizability on a completely different set of slides (Fig. 13). We found that all three models generalized well with only slight decreases in accuracy, and CLAM achieved higher performance than the MCL models. Notably, the DFCI cohort consisted of only tumor-containing slides, so we could not externally validate the performance on normal slides.

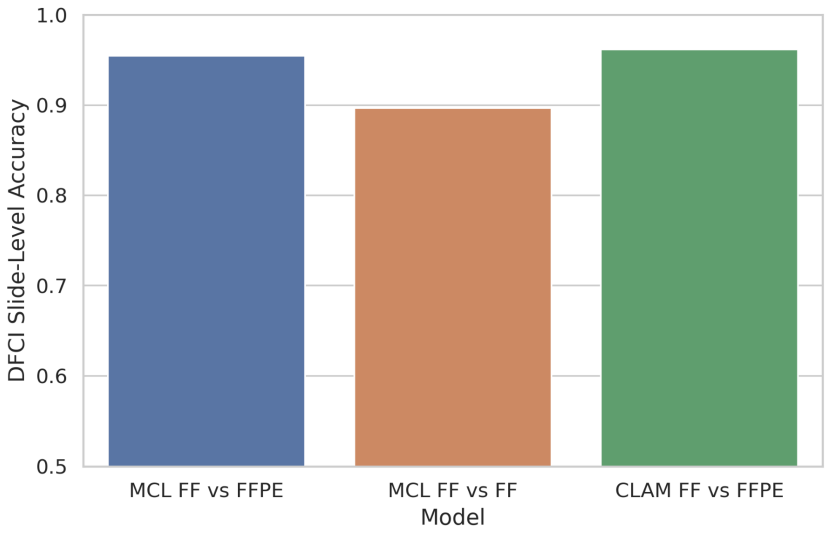

Figure 13. TCGA Tumor vs Normal Models Generalize to External DFCI Cohort. Overall slide predictions were calculated for the MCL models averaging the per-fold slide predictions across the 4 models trained during TCGA cross-validation. For the CLAM model, a new model was refitted to all TCGA training data and applied to generate slide-level predictions.

Beyond the high performance of these models on the DFCI data, we found that models trained only on the FF modality can be successfully transferred to predict the tumor/normal status for FFPE slides, although at the cost of accuracy (Fig. 13). This is a striking result given the substantial visual differences between these two imaging types, and it allows us to circumvent the lack of FFPE normal slides in TCGA by transferring an FF/FF model.

Because the DFCI data was generated at a single center that isn't included in TCGA, it may contain batch effects and artifacts that place it out of distribution relative to the TCGA data. For example, most DFCI slides contained pen markings, which were uncommon in the TCGA slides. To investigate this further, we visualized the landscape of tiles across the TCGA and DFCI datasets by applying the uniform manifold approximation and projection (UMAP) algorithm to the MCL tile embeddings following training of the network (Fig. 14)[26]. We found that TCGA and DFCI FFPE tumor tiles tended to cluster away from each other, suggesting that batch effects are present. Despite this, the models trained were robust to this distribution shift, further highlighting the generalizability of these models.

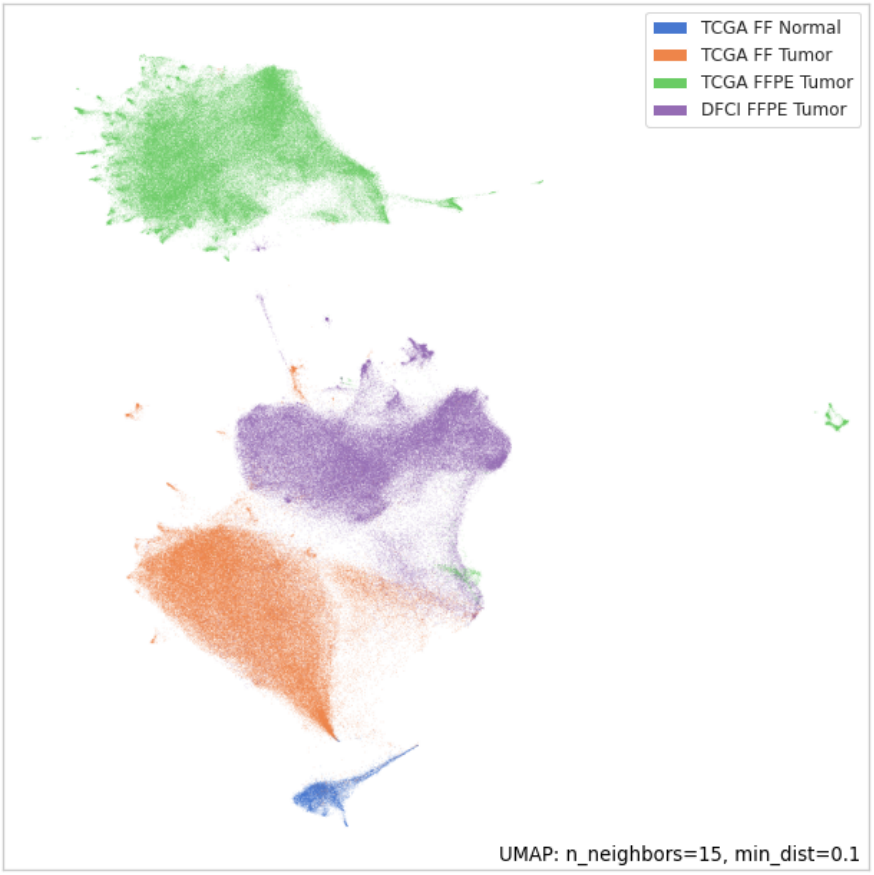

Figure 14. TCGA and DFCI Datasets are Distributionally-Shifted. FF, flash-frozen; FFPE, fixed-formalin paraffin-embedded.

2.8 Generating Slide-Level Heatmaps of Tumor

To further visualize and interpret model outputs on the DFCI slides, we applied the MCL and CLAM models to generate slide-level heatmaps highlighting tumor and normal regions of a select slide (Fig. 15, 16, 17). For the MCL models, we visualized the predicted tumor probability for each tile in the slide, while for the CLAM FF/FFPE model, we visualized the model-predicted pro-tumor attention for each tile in the slide.

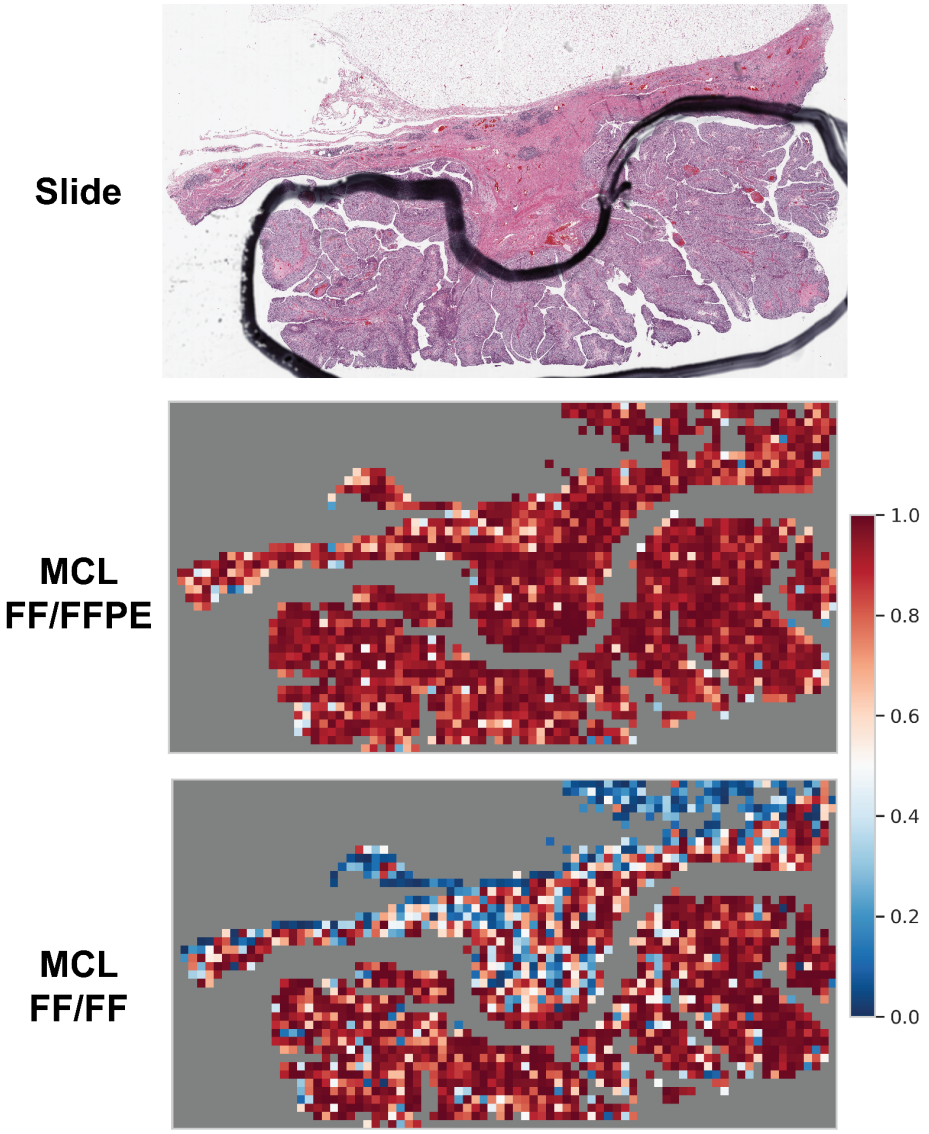

Figure 15. Transferring the MCL FF/FF Model to FFPE Slides Better Resolves Tumor Location over MCL FF/FFPE. Top: an FFPE, tumor-bearing WSI from the DFCI cohort. Heatmaps from the MCL FF/FFPE (middle) and MCL FF/FF (bottom) models were generated for this slide. Color indicates the predicted probability of each tile to be tumor by the model.

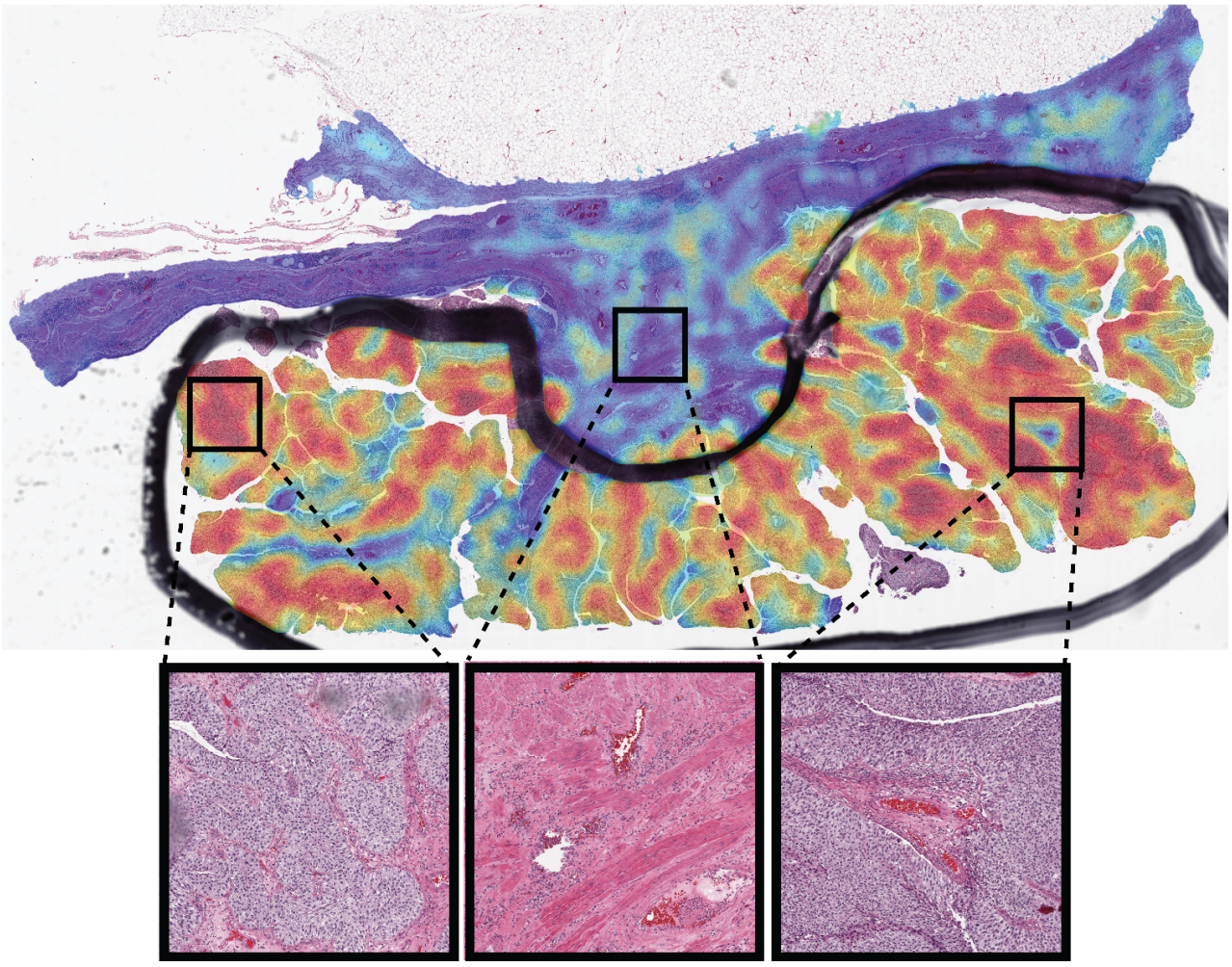

Figure 16. CLAM Heatmaps Distinguish Regions of Tumor from Stroma Within a Slide. Color indicates the predicted attention for each tile using a stride of 0.05 and smoothed with a Gaussian blur. Red indicates regions with high attention for tumor prediction, while blue indicates regions with low attention.

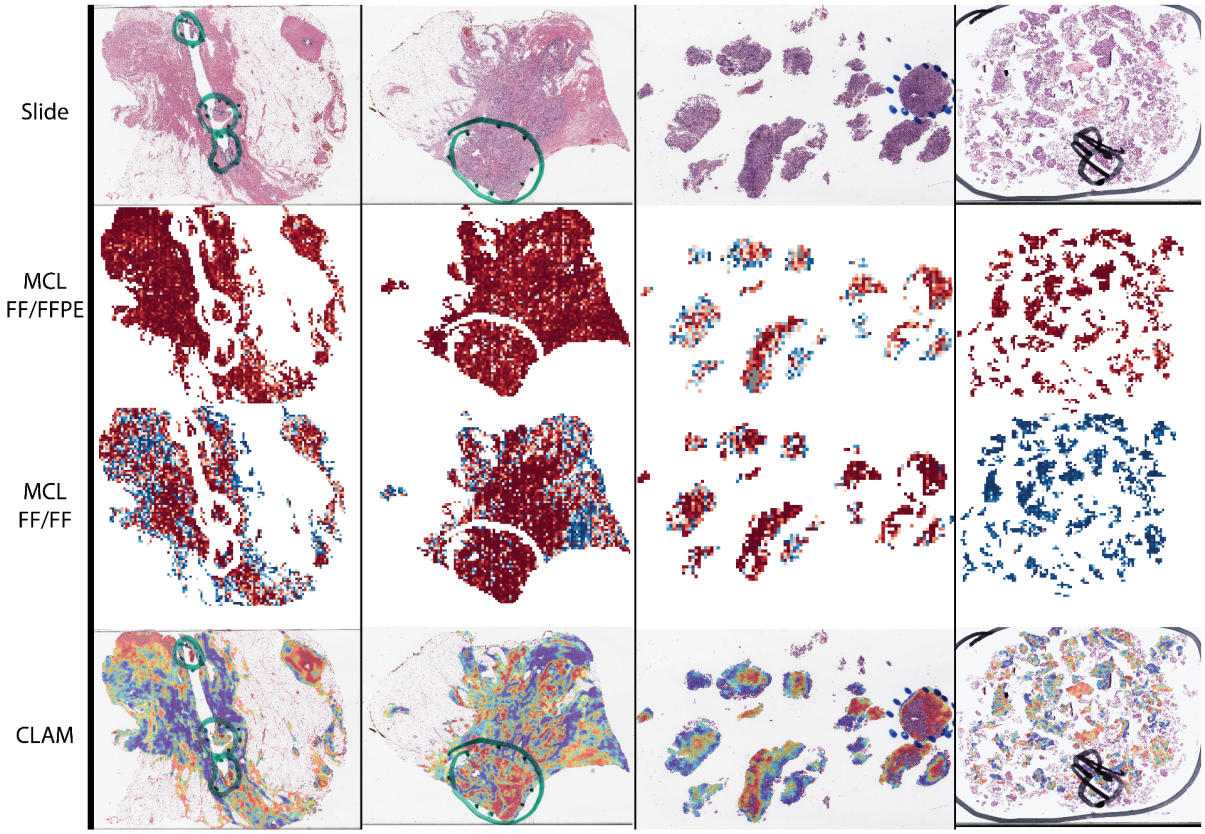

Figure 17. Comparison of MCL and CLAM heatmaps for additional slides. In column 4, CLAM heatmap scale is inverted due to incorrect prediction (i.e. red is high attention for the normal phenotype).

We found that the MCL FF/FFPE heatmaps were relatively homogenous in tumor prediction across slides, yielding little information about the spatial arrangement of the tumor tissue (Fig. 15, 17). However, transferring the MCL FF/FF model to these FFPE slides added considerable heterogeneity in prediction scores, resulting in regions of predicted-tumor and predicted-normal area across slides. For example, the stromal region outside of the pen marking in an example slide (Fig. 15) was predicted to be strongly tumor by the MCL FF/FFPE model despite being characteristic of normal tissue. Upon applying the MCL FF/FF model to this slide, however, we found much of this stromal region to be correctly predicted as normal tissue. This suggests that transferring the MCL FF/FF model to FFPE slides may further spatially resolve the tumor compared to the confounder-driven MCL FF/FFPE model.

The CLAM heatmaps, however, yielded the highest resolution spatial readout by distinctly localizing tumor and stromal regions of the slide through the heterogeneity of attention scores (Fig. 16). In the example slide shown, high-attention regions overlapped entirely with a region annotated as tumor, while low-attention regions were found everywhere else. Areas with high nuclear density and less stromal content were generally assigned higher attention, although this was not quantified. Thus, the CLAM model was able to learn a generalizable tumor vs normal classifier that could localize tumor regions on FFPE slides.

2.9 Global Visualization of CLAM-Nominated Tumor Tiles

To generalize slide-level patterns observed from the CLAM heatmaps, we visualized the latent space of tiles across all DFCI slides. We subsampled 50,000 randomly selected tiles, performed principal component analysis (PCA) on their ResNet50 embeddings, and plotted the first two principal components (Fig. 18a). By additionally binning points into a 2D grid and computing a map of average attention, we found that high attention tiles were overrepresented in one region, while low attention tiles were overrepresented in another (Fig. 18b). Representative tiles were randomly chosen from these regions, and the high-attention tiles displayed phenotypes concordant with tumor such as high nuclear density and crowding. The low-attention tiles, however, were of low quality sections of tissue, perhaps representing uninformative regions rather than pro-normal predicted regions. Notably, tiles did not cluster immediately next to other tiles from the same slide, suggesting that the latent space captures general properties of tiles rather than slide-specific artifacts. Thus, exploring the space of tile embeddings with CLAM attention reveals the high-level properties detected across all slides by the CLAM model.

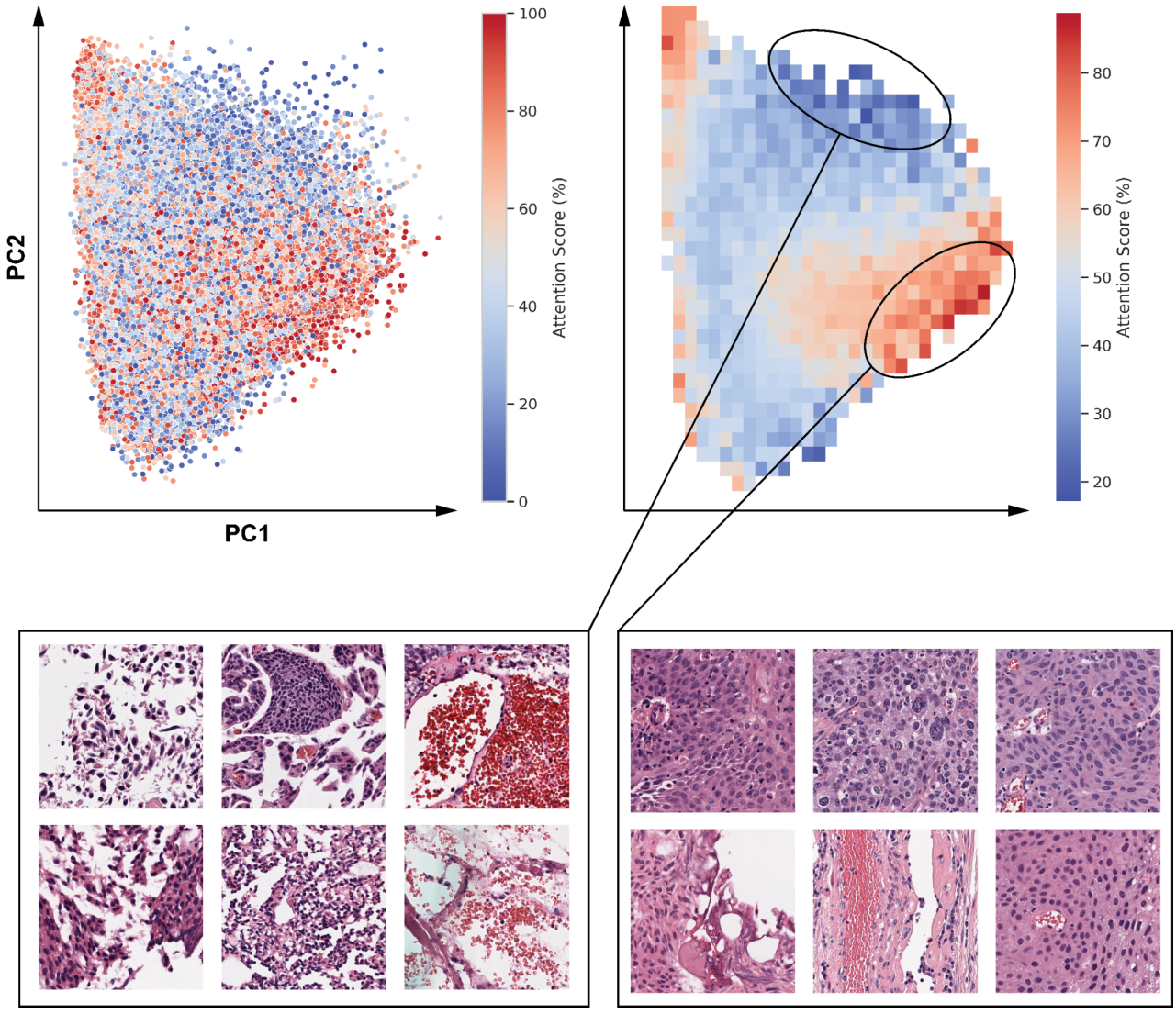

Figure 18. Tiles display shared patterns across slides. 50,000 randomly sampled tiles were reduced to two dimensions via PCA on their CLAM embeddings and visualized as (a) points and (b) 2D binning and averaging of attention.

Discussion

The prediction of tumor/normal status is a foundational digital pathology task that paves the way for more complex analyses. In this chapter, we applied the MCL and CLAM pipelines to over 1,500 urothelial carcinoma WSI's across two cohorts and two imaging modalities. We evaluated performance and generated interpretable visualizations to learn more about both the model and

the slide. In the process, we uncovered confounding factors and attempted to circumvent these by careful experimental design. Ultimately, our novel contributions in this chapter are: (1) a new nearest-neighbor approach to rigorously identifying train-test leakage, (2) showing the strong transferability of FF-trained models to the FFPE domain in the urothelial carcinoma tumor/normal context, and (3) the evaluation and visualization of the tumor/normal task against the previously unstudied DFCI urothelial carcinoma histopathology dataset.

2.10 Training and Evaluation of Models

Evaluating our model on the TCGA and DFCI datasets, we achieved high tile-level accuracy with the MCL models. For example, our MCL FF/FF model significantly outperformed a similar model trained by another group for tumor/normal prediction in urothelial carcinoma (Fig. 8) [6]. However, given that the slide tumor purity of urothelial carcinoma is expected to be lower than the tile-level accuracy of these models [6], we suspect that these models have learned confounding information present in most tiles that associate with the tumor/normal label. This is substantiated further by the relatively homogenous heatmaps these models generate, even across areas lacking malignant tissue (Fig. 10). To investigate the confounding, we used Grad-CAMs and found whitespace to be one such possible confounder - i.e. that FF tiles tend to have more whitespace than FFPE, thus confounding the FF/FFPE task. However, this does not explain the high tile accuracy of the FF/FF model and its ability to outperform the theoretically more confounded (and thus “easier”) FF/FFPE model (Fig. 8). Thus, the cause for high tile-accuracy in these models remains uncharacterized.

Strikingly, the CLAM FF/FFPE model did not encounter the same challenge with confounding signal and homogeneity of predictions that the MCL FF/FFPE model did, suggesting that a difference between these two pipelines is driving the drastically different model behaviors. One explanation is the use of a frozen ResNet in CLAM versus a fine-tuned ResNet in MCL. This may allow the MCL models to overfit to subtle patterns that aren’t captured in the CLAM embeddings. Removing the fine-tuning step to test this hypothesis may incur further challenges, as MCL’s ability to fine-tune the ResNet50 to identify new image features relevant to the task could bolster performance by increasing modeling capacity. The tradeoff between overfitting and modeling capacity here could be further studied by varying the strength of fine-tuning and observing how the MCL accuracies and heatmaps change. Another possible explanation could be that CLAM, as a multiple-instance learning model, relaxes the need for

predicting each tile correctly by instead aggregating information across the whole slide. This could encourage identification of the most salient tumor/normal features across a subset of tiles in the whole slide, rather than identifying subtle and potentially confounded features in every tile.

When comparing the accuracy of the MCL and CLAM FF/FFPE models at the slide level, we found that although CLAM performed worse than the MCL models on the TCGA test set, it performed slightly better on the external DFCI cohort (Fig. 9, 11). This may be explained by MCL's learning of some TCGA-specific confounded signal that failed to generalize to DFCI, such as FFPE/FF confounding, whereas CLAM may have been slightly more robust to this as a multiple-instance learning model that integrates slide context. Because the tumor/normal task is relatively straightforward, there is a need to apply CLAM to more complex tasks (e.g. mutation prediction), where we expect the integration of spatially separated information is necessary for improved performance.

It is important to note that this study does not provide a comprehensive comparison of the MCL and CLAM pipelines due to slight differences in some parameters between the two that may affect results. For example, MCL uses a fine-tuned 2048-length ResNet50 embedding for the tiles, while CLAM uses a frozen pre-trained 1024-length embedding. Additionally, MCL uses random crops and flips to augment the training set, while CLAM works directly with the untransformed tiles. And finally, before predicting on the DFCI test set, MCL models were retrained on ⅛ of the TCGA dataset whereas CLAM was retrained on the entire dataset due to using class-unbalanced and balanced loss functions, respectively. For a more direct comparison between these two pipelines, these differences could instead be held constant across the models while studying relative performance.

Finally, we examined the latent space across cohorts and imaging types, finding that batch effects and image modality differences gave rise to data distribution shifts that may pose challenges when transferring models (Fig. 14). In particular, DFCI FFPE tiles clustered away from TCGA FFPE tiles after dimensionality reduction by UMAP, suggesting that there are institutional differences in slide properties being captured in the embeddings that could bias model performance. However, our models proved robust to this domain shift as demonstrated earlier (Fig. 13), although characterizing it remains to be done as future work. Additionally, although the TCGA FF and DFCI FFPE tiles do not cluster together, the MCL FF/FF experiment was able to generalize to the DFCI FFPE slides. This ability to transfer from FF to FFPE slides is particularly surprising when considering the procedural and visual differences between the two modalities (Fig. 1)[15]. Thus, these results suggest that although dataset shift is certainly present, the models learned could generalize across this shift.

2.11 CLAM Heatmaps and Latent Space Analysis

The CLAM high-resolution tumor heatmaps generated here are information dense readouts for whole-slide images, akin to staining a slide for a tumor marker via immunofluorescence. The complexity apparent in the heatmaps is particularly impressive because the CLAM model was trained entirely using weak labels. Extrapolating from a slide label to precise spatial localization is known as *virtual staining* and represents perhaps CLAM's strongest use case in the digital pathology field, although visual validation of spatial outputs is necessary. Besides manual pathologist verification, virtual staining can also be validated where immunofluorescence (IF) data is collected in addition to the H&E stain. This has been used to verify virtual mRNA expression stains[27]. For the bladder cancer tumor/normal context, a stain for a tumor marker could allow for more fine-grained evaluation of the CLAM heatmaps. With paired H&E-IF datasets now being generated, this could be a powerful future approach to rigorously benchmarking the quality of CLAM virtual staining.

While the CLAM heatmaps highlight diagnostic imaging patterns in individual slides, we also interpreted the pan-cohort context by performing PCA on CLAM tile-level embeddings from all DFCI slides and overlaying the assigned attention scores (Fig. 18). We found that two regions of the latent space were enriched for high and low attention tiles separately and harbored neighborhoods of morphologically-similar tiles, indicating that the CLAM attention scores reflect visual characteristics of the tiles that are preserved across slides. The lack of tight clustering of tiles from the same slide suggests that this latent space captures robust axes of variation that integrate tiles across slides together. Using this space as a canvas, we could derive further insight across the landscape of tiles by overlaying other virtually staining properties, such as presence of tumor-infiltrated lymphocytes or microvessels. Thus, the global tumor/normal visualization here can be extended to study other tasks where the diagnostic image phenotypes may be less known.

Going forward, these approaches may also be incorporated into subsequent efforts to dissect the spatial heterogeneity of tumors, and relate these properties to molecular and clinical states in bladder cancer. These include predicting subtypes of bladder cancer harboring DNA repair deficiencies that may have therapeutic relevance (e.g. ERCC2 mutations[28]), and likewise mapping tumor and immune cell states in heterogeneous bladder cancer and their relationship to selective chemotherapy or immunotherapy response. Finally, these strategies may be broadly relevant for analyses in other solid tumors given the ubiquity of diagnostic histopathology images in cancer.

Bibliography

1. Global cancer data by country | World Cancer Research Fund International. *WCRF International* https://www.wcrf.org/cancer-trends/global-cancer-data-by-country/.
2. de Haan, K. *et al.* Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **12**, 4884 (2021).
3. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis | Nature Cancer. https://www.nature.com/articles/s43018-020-0085-8.
4. Xu, Z. *et al.* Deep learning predicts chromosomal instability from histopathology images. *iScience* **24**, 102394 (2021).
5. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
6. Noorbakhsh, J. *et al.* Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images | Nature Communications. https://www.nature.com/articles/s41467-020-20030-5.
7. Woerl, A.-C. *et al.* Deep Learning Predicts Molecular Subtype of Muscle-invasive Bladder Cancer from Conventional Histopathological Slides. *Eur. Urol.* **78**, 256–264 (2020).
8. Zhang, Z. *et al.* Pathologist-level interpretable whole-slide cancer diagnosis with deep learning | Nature Machine Intelligence. https://www.nature.com/articles/s42256-019-0052-1.
9. Saltz, J. *et al.* Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep.* **23**, 181-193.e7 (2018).
10. Berdik, C. Unlocking bladder cancer. *Nature* **551**, S34–S35 (2017).
11. Robertson, A. G. *et al.* Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171**, 540-556.e25 (2017).
12. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the

human bladder. *Science* **370**, 75–82 (2020).

13. Liu, D. *et al.* Mutational patterns in chemotherapy resistant muscle-invasive bladder cancer. *Nat. Commun.* **8**, 2193 (2017).

14. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).

15. Gao, X. H. *et al.* Comparison of Fresh Frozen Tissue With Formalin-Fixed Paraffin-Embedded Tissue for Mutation Analysis Using a Multi-Gene Panel in Patients With Colorectal Cancer. *Front. Oncol.* **10**, 310 (2020).

16. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

17. Taqi, S. A., Sami, S. A., Sami, L. B. & Zaki, S. A. A review of artifacts in histopathology. *J. Oral Maxillofac. Pathol. JOMFP* **22**, 279 (2018).

18. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clin. Cancer Inform.* (2019).

19. Nyman, J. & Narayanan, S. *MC Lightning*. (2022).

20. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015).

21. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).

22. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* (eds. Pereira, F., Burges, C. J., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).

23. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).

doi:10.1109/CVPR.2009.5206848.

24. Biewald, L. Experiment Tracking with Weights and Biases. (2020).

25. Hari, S. N. *et al. Examining Batch Effect in Histopathology as a Distributionally Robust Optimization Problem*. 2021.09.14.460365 https://www.biorxiv.org/content/10.1101/2021.09.14.460365v1 (2021) doi:10.1101/2021.09.14.460365.

26. Sainburg, T., McInnes, L. & Gentner, T. Parametric UMAP Embeddings for Representation and Semisupervised Learning | Neural Computation | MIT Press. https://direct.mit.edu/neco/article/33/11/2881/107068/Parametric-UMAP-Embeddings-for-Representation-and.

27. Schmauch, B. *et al.* A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.* **11**, 3877 (2020).

28. Van Allen, E. M. *et al.* Somatic ERCC2 Mutations Correlate with Cisplatin Sensitivity in Muscle-Invasive Urothelial Carcinoma. *Cancer Discov.* **4**, 1140–1153 (2014).